DOCUMENT RESUME

ED 305 721 EA 020 804

AUTHOR Kirst, Michael W.; And Others
TITLE Evaluating State Education Reforms: A Special Legislative Report.
INSTITUTION National Conference of State Legislatures, Washington, D.C.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 86
CONTRACT 400-85-1004
NOTE 134p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Information Analyses (070)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Change Strategies; *Educational Change; *Educational Policy; Elementary Secondary Education; *Excellence in Education; *Instructional Improvement; *Personnel Evaluation; Program Effectiveness; State Legislation; Teaching Methods

ABSTRACT

Evaluating the impact of the recent state education reforms is an issue now confronting every state legislature. This paper focuses upon evaluating state education policies for a legislative audience. The text of the paper covers the following topics: (1) the urgency of staff evaluation; (2) how evaluation can help state legislators; (3) reasonable expectations for evaluation; (4) the use of tests in measuring the effectiveness of educational programs, methods, and instructional methods; (5) new directions for state education data systems; (6) state and local performance indicators and cost/effectiveness studies; and (7) how well U.S. schools perform. (SI)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED305721

# EVALUATING STATE EDUCATION REFORMS:

## A Special Legislative Report

Michael W. Kirst

Stanford University

With contributions by:

Ralph W. Tyler
Kimberly Ford
John Rogers

Prepared for the National Conference of State Legislatures and funded by the U.S Office of Educational Research & Improvement
1986.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy



BEST COPY AVAILABLE

EA020804

This special report was funded by a contract between NCSL and the Office of Educational Research and Improvement, U.S. Department of Education (Contract #400-85-1004.) However, the contents do not necessarily reflect the position of the Department nor any of its agencies.

## PREFACE

Evaluating the impact of the recent state education reforms is an issue now confronting every state legislature. Are the reforms being implemented? Are they effective in improving teaching and learning? What are reasonable expectations in upgrading student achievement? How can lawmakers sustain the public's commitment -- and their own -- for improving education over time?

Because of the widespread interest in these questions, the National Conference of State Legislatures sought to provide some answers. We therefore asked Dr. Michael Kirst to write a document that pulled together the "state of the art" on evaluating state education policies for a legislative audience. As a Professor of Education at Stanford University and a former President of the California State Board of Education, Dr. Kirst was a logical choice. As a prolific writer on the politics of education and as co-director of PACE, a state education policy research group currently evaluating California's 1984 education reforms, he was, we feel, the best choice.

NCSL wishes to thank Dr. Kirst for producing this document for legislators and their staffs. His publication should prove extremely useful as they wrestle with the thorny process of evaluating recent state efforts to improve the schools. We are also deeply grateful to the U.S. Office of Educational Research and Improvement, without whose support this document could not have been written.

Peggy M. Siegel
Education Program Manager
National Conference of State Legislatures

The National Conference of State Legislatures is the official representative of the nation's 7,461 state legislators and their staffs.

NCSL was created in January 1975 from the merger of three organizations that served or represented state legislatures. NCSL is a nonpartisan organization with three objectives:

- To improve the quality and effectiveness of state legislatures;
- To foster interstate communication and cooperation; and
- To ensure states a strong, cohesive voice in the federal system.

Earl Mackey
Executive Director

The Conference operates from an office in Denver, Colorado and an office in Washington, D.C.

## ACKNOWLEDGMENTS

This handbook in several ways is a joint effort of the Policy Analysis for California Education (PACE) research group at Stanford University. PACE analyzes state policy issues and the condition of education in California. Three advanced graduate students (John Rogers, Kim Ford, and Worku Negash) worked on the third chapter. Ford and Negash were particularly helpful concerning the evolution of program evaluation. Rogers authored the cost-effectiveness part of chapter 6.

Peggy Siegel and Bill Harrison of NCSL in Washington provided a thorough critique of two drafts. They were helpful in all phases of this endeavor.

The Stanford Alumni Association and W. H. Freeman granted permission to reprint "How Well Do Our Schools Perform?". This was originally Chapter 5 in my book, Who Controls Our Schools? (New York: W. H. Freeman, 1984). The McCutchan Publishing Corporation of Berkeley, California, permitted the use of Ralph Tyler's chapter on testing from their book, Crucial Issues in Testing, by Ralph W. Tyler and Richard M. Wolf. Sage Publications permitted the use of some excerpts that I wrote for their May 1986 issue of Education and Urban Society.

Michael W. Kirst

Stanford, California, 1986

## TABLE OF CONTENTS


| | | |
|---|---|---|
| Chapter 1 | The Urgency of State Evaluation | Michael W. Kirst |
| Chapter 2 | How Evaluation Can Help State Legislators | Michael W. Kirst |
| Chapter 3 | Reasonable Expectations for Evaluation | Michael W. Kirst, Kimberly Ford and Worku Negash |
| Chapter 4 | The Use of Tests in Measuring the Effectiveness of Educational Programs, Methods, and Instructional Materials | |
| Chapter 5 | New Directions for State Education Data System | Michael W. Kirst |
| Chapter 6 | State and Local Performance Indicators and Cost/Effectiveness Studies | Michael W. Kirst and John Rogers |
| Chapter 7 | How Well Do U.S. Schools Perform | Michael W. Kirst |

## Chapter 1

## THE URGENCY OF STATE EVALUATION

**Michael W. Kirst**

Concern about the quality of American education has virtually exploded in the last year and a half. There may or may not be a rising tide of mediocrity, but states certainly are riding the crest of a wave of initiatives in public education.

This brief period of national interest in education -- which probably will last no more than five years -- has been characterized by an intense policy response on the part of most states. The Education Commission of the States (ECS), for example, reported that as many as 290 high-level state commissions were studying the quality of public education in 1984. It was not unusual for the governor, state legislature, and chief state school officer to have their own separate commissions. And these commissions have been responsible for a great deal of new state policy.

Many states already were considering major "reform" legislation before the 1983-85 spate of highly publicized national reports on education. Other states did not make major moves until 1986. This recent wave of reform, however, is a continuation of the aggressive state initiatives of the past 20 years. The growth in state oversight of local education is the most striking feature of state-local relations in the last two decades. From 1983 to 1986, the organizations of professional educators and local schools were making suggestions for only marginal changes in these state omnibus bills. And under the Reagan administration, the federal role has been mostly restricted to small pilot programs.

These trends promise a basic restructuring of federal, state, and local relations that will cede considerably more education initiative to states. This new state-local relationship, however, will include an enormous range of

variation in state policy aggressiveness -- from the highly prescriptive states, such as California and Florida, to the more local control-oriented states, such as Colorado and New Hampshire.

The 1983-86 spurt in state activity intensifies the steady growth in state policy throughout the 1970s, when states became involved in school finance reform, categorical programs, school improvement efforts, minimum competency testing, and civil rights regulations, for example. But is was not until the early 1980s that state governments provided the majority of the current operating funds for education.

Dangers, however, attend aggressive, broad-based state action that changes policy through statutes and regulations, which have a standardizing effect. Moreover, the new focus of state policymaking is no longer on particular targets, such as the handicapped or minority students; instead, it is aimed at the core of basic instructional policy, including what should be taught, how it should be taught, and who should teach it.

The crucial policy question for the next five years is whether this state "reform momentum" will keep going. If it does, then education expenditures will exceed inflation, and underlying negative fiscal trends will be pushed to the background. However, if the public and key policymakers perceive that education reform has failed or has not been implemented, then a less optimistic future for education is likely. If the public perceives that professional educators have subverted the aims of state reformers, profound disillusionment and loss of confidence will ensue. Consequently, implementation and assessment of these reforms should be a top priority for educational research and state policymakers. In 1986, the pace of new reforms has slowed significantly as fewer states and local school districts enacted new omnibus bills or major interventions. Moreover, national economic growth is slowing, and the decline in commodity prices is lowering state revenues, setting up tough competition for increased state funds. Many of the state initiatives were untried and untested concepts such as teacher career ladders and statewide testing of veteran teachers for recertification. Evaluating the impact and effectiveness of the

varied education reforms therefore becomes urgent because not all of them can be expanded or even maintained.

Education policy is now beyond the 1983-84 "alarmed discovery" and "crisis activity" phases of the issue-attention cycle. Other reforms such as ecology and urban development, have degenerated into the subsequent policy phases of "disillusionment with results" and then a "return of neglect.[2] The process of implementation, adaptation, refinement, and elimination of the state education reforms has begun in earnest. For example, a widespread realization exists that teacher quality is crucial to increased academic attainment, but states are unsure how to blend screens and magnets to attract and retain high-quality teachers. Consequently, states are trying all kinds of teacher quality interventions, including career ladders, minimum salaries, mini-grants, sabbaticals, and forgivable loans, without a clear notion of which approaches are optimal. It is doubtful that many states can afford to fully fund all of the various proposals to improve the teaching profession, so evaluation must help policymakers decide which are their best bets.

Sustaining the reform momentum becomes even more urgent when considering that national enrollments will grow by 2.1 million by 1990. Odden estimates that it will take 5 percent annual real growth in total education revenues merely to pay for this enrollment increase![3] He emphasizes that total revenue grew only 7 percent after inflation between 1980-85. The reforms recommended by the National Commission on Educational Excellence would require about a 20 percent annual increase in per pupil expenditures. Only two out of eight reform states that Odden studied have approached this level of increase. Since increases in state aid of this magnitude are unlikely, policymakers must sort out which of the many state reforms should be expanded, eliminated, or left at their current funding level.

It is now time to advance some working definitions for the terms evaluation and policy analysis. Evaluation means the process of ascertaining or judging to what extent the implemented activities and procedures of local school districts have met the predetermined objectives of the state education reforms. Although

evaluation is retrospective, it provides important insights into policymakings. Policy analysis is the process of making theoretical or empirical forecasts of the extent to which a plan of state activities and procedures will meet the state's educational reform objectives. In this way, policy analysis is prospective rather than retrospective. Given the status of reforms from 1983 to 1986, state policymakers will increasingly turn to retrospective analysis before proceeding with many bold new programs.

Ultimately, state policymakers will have to judge how well they perceive that state reforms. In short, no single, simple answer exists that policymakers can use. The notion of one indicator like "runs batted in" or "miles per gallon" as a measure of these complex and ambitious state reform programs is a misconception.

Hence, it appears that evaluation approaches should be multiple and should explore the local school district from many different directions. Moreover, the best evaluation strategy is sequential, starting with whether a program has even attracted the attention of local authorities, and later moving to whether they have made any move to implement or institutionalize the state's intent. Next, policymakers should examine the differential impact of the same state reform program in diverse local school districts.

Many urgent policy and strategy questions are subsumed within the state education reforms. What, for example, is the proper balance between state and local control? Will bottom-up commitment at the school site reinforce top-down state leadership and lead to a state-local partnership toward the same goals? While most school districts are increasing required academic courses, what will make all students motivated and interested in these added courses? Do some reforms, such as merit pay, outrun our current technology to implement them effectively? As science and match enrollments increase dramatically, who will teach these courses? What will these reforms do to the dropout rate?

In 1986, the 50 states provide an unprecedented natural setting to answer these and other important questions. The major state policy issues of the next

five years will focus on state education reform problems, successes, and unanswered questions. State need to involve scholars and practitioners from a variety of perspectives to help answer these complex questions. Fortunately, enough diversity exists in state policy approaches and local contexts to provide a quasi-natural experiment nationwide. Some states in initiating teacher reforms, for example, have imposed a statewide career ladder model, while others have relied on locally generated differences in approaching career structures. Evaluating all the policy issues raised in the 1983-86 state reforms will be expensive, but only a few states, such as California, South Carolina and Tennessee, have earmarked significant money for in depth analysis of the impact or success of these reforms. States often use an audit approach when a more thorough evaluation, as outlined in this report, is necessary. It is ironic that although education has so much at stake concerning the public's perception of these state reforms, so few resources are devoted to assessing their outcomes. The 1983-86 wave of state reform is the most ambitious policy change many states have ever undertaken. Consequently, informed judgements that give these reforms a chance to work are urgently needed.

The recent explosion of state legislation has led to new questions about how to evaluate this large-scale and diverse educational innovation enterprise. However, evaluation is only part of the information needed to make the decisions just outlined. It can enhance legislative judgement, but not replace it.

## THE POTENTIAL OF EVALUATION

In its broadest sense, evaluation can assist in many important state decisions. For instance, legislators and other state policymakers must decide which of the many claimants for scarce education funds should receive increases. Should newer programs receive priority over more established ones? What provisions should be altered in the most recent reform bills? Which initiatives are not being implemented or appear to be misguided? Which pilot programs merit statewide expansion? What provisions of the law could cause unintended and unanticipated consequences? When should the legislature loosen a mandate for

local schools? Which reforms are embraced enthusiastically by local educators but are seriously underfunded? What early or midcourse changes can help the reforms meet their potential? What are the appropriate time and circumstances for deciding whether a particular program works?

A persistent theme throughout this report is that the commonsense initial reaction about evaluating reform does not always work. Indeed, many states confuse auditing with evaluation and rarely establish an evaluation strategy and program soon after beginning a major initiative. Auditors and accountants with little background in classroom instruction are sometimes sent to school sites with little guidance beyond looking for expense vouchers and budget documents. Frequently it takes budget cutbacks to make policymakers consider evaluation, and then, only to determine what programs to cut. But many times, state policymakers may contract with outside consultants to conduct the evaluations or they may expect state education departments to perform this function. Often, however, the state education department has not been provided adequate resources for evaluation and also may lack credibility with some legislators because it administers the program to be evaluated, or the legislature may realize the need for evaluation early in the cycle but may want definitive answers too quickly. Consequently, policymakers may be tempted to pull up the plant by its roots in order to see if it is growing.

This special report provides an overview of concepts, lessons from the past, and guidance for state policymakers who desire to evaluate the state education reforms enacted from 1983-86. It is not a technical manual for experts, nor does it purport to be comprehensive. The objective is to help state policymakers think about the uses, limits, and strategies of evaluations. Legislators will not conduct the evaluations, But they can benefit from understanding basic concepts, guidelines, and mechanisms for commissioning and implementing evaluation, and making decisions after receiving them.

Chapter 2 provides an overview of the major evaluation techniques and concepts. It is the key integrative chapter of this special report and refers

the reader to the other chapters for detailed discussions of specific topics. It can be read as an executive summary for the entire report.

After Chapter 2 analyzes what legislators can do with evaluation, Chapter 3 provides some caveats and complexities about oversimplified uses and expectations from evaluation. Chapter 4 continues this theme with a detailed examination of the uses, limits, and misuses of tests.

Chapter 5 examines the drastic changes in the basic state data bases that are needed to begin evaluating the recent reforms. Current state data are oriented to an older agenda of categorical programs and school finance distribution formulas.

Chapter 6 provides more depth on a major emerging evaluation issue of performance indicators. In November 1985, the Council of Chief State School Officers voted to develop interstate indicators of tests, course enrollments, social context variables, and other indicators. Individual states also are constructing such barometers in order to gauge some trends. Performance indicator systems are in their infancy and merit careful attention for developing more sophisticated measurement techniques.

Chapter 6 also provides information on an essential, but often overlooked evaluation dimension -- cost studies. Most states did not carefully cost out each of their recent education reform initiatives; rather, they developed a list of initiatives and then appropriated an overall amount for all or most of them. Often the total appropriated was determined by the current status of the state treasury rather than by a detailed analysis of the cost of essential ingredients in each reform. Chapter 6 combines cost analysis with the outcome dimension through a general description of cost-effectiveness techniques. This will help state policymakers decide whether they want to commission cost-effectiveness studies.

Chapter 7 is designed to provide state policymakers with an overall grasp of how well the U.S. education system is doing. This chapter emphasizes international comparisons and long-term trends. It highlights the basic

objectives of U.S. education and how they differ from those of other nations. Media reports comparing the education systems of different countries have frequently drawn simplistic conclusions, leading to considerable confusion. A more realistic perspective is essential to assess the performance of America's system in a fair manner, and decide what level of state intervention in local school operations is warranted by international considerations.

## Chapter 2

## HOW EVALUATION CAN HELP STATE LEGISLATORS

Michael W. Kirst

State education reforms should be evaluated from several different conceptual approaches. The choice a state makes depends partly on such factors as: (a) the quality of its existing and future data base, (b) the amount of money it can spend, (c) the quantity and quality of trained evaluation personnel, and (d) the specific type of reform package. After presenting seven different types of evaluation approaches, we will return to the mix and choices that might be appropriate for a specific state. States with omnibus bills might need to use most or all of these evaluation approaches. Other states may need only a few of them for the major discrete programs in a more limited reform bill.

The next section presents several major types of evaluation that a legislature might use. There is a separate chapter that follows up on many of these seven approaches. The purpose here is to introduce the major approaches and refer the reader to the appropriate chapter in this report for more detail. These seven approaches are the organizing rubrics for thinking about types of evaluation. The mechanisms for evaluation such as use of legislative oversight committees follow.

1. Performance indicators. Performance indicators are statewide numerical measures of trends in educational variables. In some cases these standardized state measures can be supplemented by locally devised indicators that vary in definition and concept depending on local conditions. Performance indicators pick up changes that are easily measured, but they can rarely penetrate behind the classroom door to measure such things as the content actually taught, teacher morale, or the type of intellectual tasks students are performing. Performance indicators are useful parts of a statewide assessment strategy, but

if used alone they tend to overvalue what can be measured at the state level. Some states have created a system of statewide targets and individual school profiles for numerous uniform indicators, including changes in course-enrollment patterns, test scores, number of dropouts, and performance of college freshmen. Statewide indicators require a good state data base as discussed in Chapter 5. These statewide indicators are supplemented by locally devised indicators in area not easily measured at the state level, such as school climate, time spent on writing, and amount of homework. States have used public opinion surveys of teachers, pupils, and parents concerning their views on the reform impact and design. We discuss these performance indicators in detail in Chapter 6.

2. Overall studies of the financial impact of reforms. Rather than carefully compare the cost of each reform, states have backed into allocating funds to the local districts according to the amount of uncommitted state revenue. Often money is provided through a state's basic finance formula, and the money is not tied to any specific reforms. Consequently, states need to know where the districts have spent the increased funding and whether some components of the reform effort have been over or underfunded. For example, states need to know how much money was actually required to institute state-mandated science courses or to create new approaches to local teacher evaluation. They also need to know which areas of the curriculum gained by the funding increases (usually math and science) and which areas lost (often home economics and industrial arts). States also need to know whether the reforms become more or less expensive as they become part of collective bargaining and whether local use of money varies according to the prior spending level of a district.

In order for a reform to have any chance of working, some new resources must usually be provided. Cost analysis will help determine how much was spent in what areas, but it will not tell us anything about the results. We discuss cost analysis in more detail in Chapter 6.

3. Analysis of cost-effectiveness of various state interventions with the same specific objectives. Henery Levin states the case for cost-effectiveness this way:

> (Cost-effectiveness) integrates the results of (program) costs in such a way that one can select the best educational results for any given level of educational results for least cost. It is important to emphasize that both the cost and effectiveness aspects are important and must be integrated. Just as evaluators often consider only the effects of a particular alternative or intervention, administrators sometimes consider only cost. In both cases, the evaluation will be incomplete.

Cost-effectiveness can provide important policy information, but it is limited to comparisons among programs with similar objectives. A possible example would be the use of loans, scholarships, or higher base salaries as a magnet to attract better-quality beginning teachers. Chapter 6 focusses on cost-effectiveness.

4. Program evaluation. Some states have created programs that can be evaluated as discrete activities, such as career ladders, preschool programs, or increases in the number of high school counselors. These programs can be evaluated by means of well-developed techniques of program evaluation that have been used to assess such federal programs as Chapter I. As a first step, program evaluation can research what components of a discrete program were actually implemented. It can then move to include costs, outcomes, and processes. A comparison of several programs with very similar objectives could constitute a study of their comparative cost-effectiveness.

Program evaluation that focuses on separate initiatives cannot address well the interaction of several different state initiatives, however. Nor can it give us much insight into the cumulative impact of omnibus state reform activities. Program evaluation, however, can enlighten us on successful patterns of local implementation impediments to program success and on what state policies help implementation.

A program evaluation of Chapter I of ECIA was funded in 1984 by the U.S. Congress. This is an unusually large undertaking but does demonstrate most

components of a complete program evaluation. The topics covered by the various program evaluators include:[1]

1. The nature and extent of students' need for compensatory education services -- this includes the relationship between poverty and school achievement as well as changing demographics. The program targets compensatory education services to students that receive special services on top of the regular school program.

2. The current operation of the program and the prospects for improving it.

   a. The costs and contributions of program administration, including the distribution of administrative responsibilities across levels of government.

   b. Targeting funds to the most needy schools and students. How do districts determine which schools and students receive services?

   c. The processes and rationale for program design decisions including curriculum and non-instructional services.

   d. The character and quality of the services actually delivered to children including training of staff, coordination with other programs, and educational experiences.

3. The size and variability of program effects -- including several definitions and criteria for measuring the program's effects, such as basic and complex academic skills, school attendance, and pupil attitudes.

The study is being conducted in a nationally representative study of school districts and will take two years to complete. Note the intent is to isolate the target students and services provided by the federal compensatory education from all the other school resources and educational services that the child receives. This is a crucial objective of program evaluation, because it carves out the operation of a specific governmental intervention from the overall array

of school operations. If this can be done clearly then a program evaluation can proceed. But even in this kind of discrete program situation, evaluations very rarely reach overall summative conclusions as to whether the program is "working" or is a "success."

This type of program evaluation, however, can guide legislators in deciding policies such as:

1) Which children should receive services, e.g. those who are low in achievement or those living in extreme poverty;

2) Program modifications that will improve effectiveness, e.g., the use of programs that pull children out of classrooms or place teacher aides in the classroom;

3) how to cut down on administrative costs;

4) the distribution of responsibility between central governments and school-based decision makers (states or local central offices vs. shcool sites);

5) an overall sense of whether the program merits expansion or should have a stable budget until administrative problems are ironed out;

6) modifications to eliminate use of funds by local entities that do not meet state objectives.

5. Impact of several state interventions with the same general goal. It is probably still premature to ask whether state reform is working in terms of student achievement. It is sensible, however, to explore whether or not a reform has been implemented and, if so, whether its implementation is consistent with the broad objectives of state policymakers. Numerous state policies are directed toward the goal of increasing the "rigor" or "challenge" in the high school curriculum. The primary strategy is to increase the time students spend studying traditional academic subjects. Minimum state graduation requirements, tougher college entrance requirements, model state curricula, and the addition

of science and social studies sections to statewide tests are some examples of policies designed to make secondary education more rigorous.

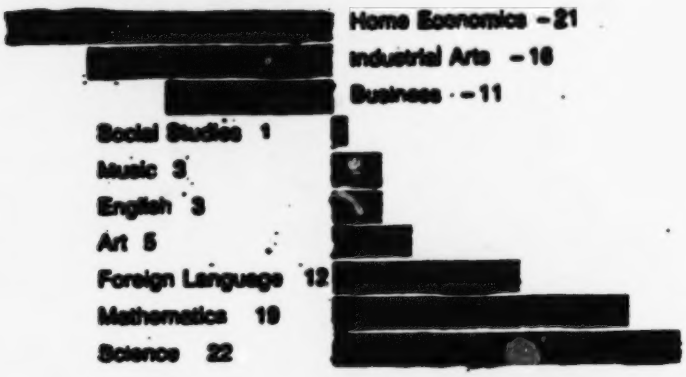
An example of an evaluation of this type of reform is a recent California study that examined changes in high school course offerings between 1982 and 1985. A example of secondary schools was surveyed, and numbers of class sections in each departmental area were taken from teachers' master schedules. After adjusting for changes in enrollments, it became clear that substantially more sections of mathematics, science, and foreign languages were being offered, while the numbers of courses in home economics, industrial arts, and business were decreasing (see Figure 1). In science, the largest increase in offerings occurred in the physical sciences, apparently in response to the new graduation requirement of one year of physical science. All areas of mathematics increased, but computer science showed the largest increase (91%), followed by more advanced math course, such as calculus, geometry, analytic geometry, and trigonometry. Calculus and analytic geometry were offered 33% more often, while general math courses increased by only 11%.

The same study showed an even more substantial increase in the number of advanced placement course offerings. By 1984-85 the number of such courses offered in chemistry, physics, and European history had increased by 34%. In addition, new advanced placement courses were offered in computer science, foreign languages, art history, and music.

An obvious problem with these simple impact analyses is that there is no way to demonstrate cause-and-effect relationships. For instance, changes could be caused by local school board policies, by state interventions, or by other factors. But if the direction of change is toward more academic coursework, state policymakers will be interested -- even without a precise analysis of the unique state role or of the content covered in these courses.

6. <u>Studies of the cumulative effects of all state reforms in omnibus bills</u>. The total number of initatives in many states makes it impossible to conceive of reform as a discrete program, such as Head Start, or a discrete

Figure 1
Percent Change in Number of Course Sections Offered, Adjusted for Changes in Enrollment (1982-83 to 1984-85)




Source: Pam Grossman, Michael W. Kirst, Worku Negash, and Jackie Schmidt-Posner, Study of Curricular Change in California Comprehensive High Schools; 1982-83 to 1984-85 (Berkeley: Policy Analysis for California Education, Policy Paper No. PP85-7-4, July 1985). Course sections were cross-checked with course enrollment data derived from state reports.

policy, such as a civil rights mandate. Instead, reforms in South Carolina and Texas, for example, contain:

- o Broad, multiple targets. Reform packages seek to alter several components of school and district operations simultaneously.
- o Lack of programmatic articulation. In several states many kinds of reform elements have been combined under a single statute. Education reform is a set of minimally related policies that will generate various responses at the local level. Given the nature of these state reforms, Michael Knapp and Marian Stearns argue the the evaluator should study the local system, not the state program. Specifically, they contend that school efforts will:

> derive from many small changes that cumulatively shift the climate for education, the perceived opportunities, the tenor of the curriculum. These shifts will be best detected by "taking the pulse" of the local educational system in ways that capture more than conventional indicators such as student test scores. . . . The many pieces of the reform agenda complete with a buzzing universe at the school and district levels for the attention of educators (and their relevant local consitituencies); second, provide positive incentives for committing further energy to education (by current staff, as well as by new recruits); and third, generate hope for, and supportive imagery of, the schools among students, educators, and the public. Accordingly, evaluation research must document what is (and isn't) noticed at the local level, and determine the effects reform initiatives have on local motivation and morale (at the administrative, teacher, and student levels). In such reform movements the whole is greater, and far more important, than the parts. Those aspects of the local scene that reflect the whole -- such as the commitment educators feel to form goals -- are consequently the most appropriate indicators oof reform efforts.[2]

In short, local case studies and state-level interviews could be used to determine the level of commitment of local actors to the reform objectives and the adequacy of available resources to support local efforts.

In conducting this kind of evaluation, it is essential to differentiate initial from long-term impacts and to be aware of changes in local response over time. The initial effects will be evident in inputs and processes, such as

changes in course enrollment patterns. Only after several years, however, should we expect significant changes in such areas as standardized test scores. This type of research is extremely complex and expensive, and it takes a long time before even the initial results can be reported. It requires longitudinal, in depth case studies of a sample of local schools within each state. Recent research on school effectiveness and school improvement shares some characteristics with investigating cumulative effects.

7. Research that isolates cause-and-effect relationships. State policy makers need to know whether there are identifiable cause-and-effect relationships between student achievement and such interventions as state-mandated curriculum alignment. Finding out will be difficult and expensive. How can we separate the effects of local policies from those of state policies when they either reinforce or work against one another? For instance, such innovations as career ladders may attract better qualified candidates to the profession of teaching, while state-mandated, test-driven curriculum standards may repel the very same people. Chapter 3 focuses on this issue.

Research on cause-and-effect relationships is probably best attempted in the later stages of assessing reforms. First we should discover which programs have been implemented and whether a program has had any impact before we undertake sophisticated studies of cause and effect. There is no sense in researching cause and effect with respect to a program that never caught on with local educators.

Some experts contend that cause-and-effect or input/output studies are not appropriate for asessing state reforms. The effects of social programs cannot be proved in the same way that one can prove a geometrical theorem or confirm a principle of physics. There are often multiple causes of educational change. The best evaluation of state reforms would assess implementation in various states and in local districts and schools within them.[3] To focus first on implementation is crucial because what is delivered to children from state reforms varies greatly according to the specific local setting. Moreover,

implementation is a multi-stage developmental process whereby local educators learn and adjust as they install the reforms.

As the education reform movement matures and bears fruit, state policy makers must keep in mind that decisions made closest to children will always be the most important. Classroom teachers vary greatly in the ways in which they react to and adapt external, state-mandated reforms. And state-level leaders who wish to assess the impact of the reforms they have mandated must begin by assessing implementation at the local level. One cause may be crucial in a particular context, but another cause will be most salient in another local context. Certain state interventions may increase the probability that a local effect will occur, but they won't inevitably produce it.

Considerations for Implementing Evaluation

Before choosing among the seven approaches the legislature will need a grasp of the status and capability of the existing state data base. An indicator system, for example, takes a great deal of time and money to construct. Some states have added questions or performance items to their existing statewide test system. Others have begun new data instrucments that collect statewide sample data on such areas as course enrollments, homework patterns, and teacher qualifications. State data systems can be costly and careful estimates need to be made.

The state education agency (SEA) is a logical candidate for conducting these evaluations, but several problems arise here. Some legislators are not confident that the SEA has the objectivity to evaluate its own administration of reform. Some SEA leaders are potential political rivals to legislative leaders. A more common issue is the failure of many states to provide funds and salary levels so that the SEA can attract and keep a substantial group of professional evaluators. Indeed, it was the federal government that inspired the state offices of planning and evaluation through its funding under Title V and IVC of the Elementary and Secondary Education Act. Now that federal funds have been cut back and state administrative burdens expanded from recent reforms, it is

appropriate for the legislature to review its support of state evaluation personnel.

States, however, can also choose to contract with non-government third parties, including universities and applied research profit or non-profit organizations. These organizations can assemble a wide variety of scholars from different academic disciplines and evaluation specialties. Many of them have been conducting evaluations of large scale governmental initiatives for over 20 years.

Some states, such as Illinois, Missouri, South Carolina, Tennessee, and Texas, have chosen to rely heavily on expanded oversight mechanisms by the legislature itself. One device would be to create a special committee to review what types of information are needed, hold hearings on emerging problems and successes, and deliberate on the results from the outcomes in the seven valuation approaches decribed earlier. Such a committee could also keep abreast of the comparative six-state evaluation conducted by the center for Policy Research in Education, (CPRE) federally funded center at the Eagleton Istitute of Rutgers University. CPRE wil be using a standard evaluation design that incorporates many of the approaches covered here.

Several states have made evaluation an integral component of their reform efforts. For example, Arizona has included pilots or demonstration programs for teacher careers ladders in their statute. The evaluation could be designed to help the legislature decide on a statewide policy. Different models of teacher career ladders could be adapted to local circumstances and then evaluated in a comparative manner. The local sites also would be chosen to reflect variation of key local factors known or believed to affect differentially career ladder outcomes. For example, some districts could be very large or small. In this type of evaluation, the legislature does not end up with numerical fundings about local responses to state reform for all school districts in a state. Rather, the study produces statements about the relationship between observed local responses to a career ladder under varied local conditions and for different local conceptions of what a career ladder should be.

In 1985, the Florida Legislature was awarded a cost-sharing grant from NCSL, funded by the U.S. Department of Education, to conduct an analysis of three Florida reforms: the Master Teacher Program, Merit Schools, and the Extended School Day. A comprehensive survey was sent to each district and site visits were made to six. Florida found that "The Master Teacher Program has received an overwhelming negative reception by superintendents, administrators, and teachers."[4] The Merit Schools program was found to be a "study in contrasts" -- some counties were enthusiastic, others were opposed and wanted to terminate it. These Florida examples demonstrate the use of program evaluation and differential local implementation effects, information that became critical to ?he Legislature in completely overhauling the master teacher program during the 1986 legislative session.

State legislators may often overestimate or underestimate the potential for evaluation to provide them with answers to vexing policy issues. It is therefore appropriate, now that the reader has had an overview of the types and strategies of evaluation, to explore in depth some of the crucial components and complexities of evaluating state reform.

## Chapter 2

## NOTES

1. U.S. Department of Education, "A Plan for a National Assessment of the EICA Chapter I," unpublished, November 1984.

2. Michael Knapp and Marian Stearns, "Improving System-Wide Performance: Evaluation Research and the State Education Reform Movement," in Joe Wholey, ed., Toward Excellence: Roles for Evaluators (Lexington, Mass.: Lexington Books, 1986).

3. Milbrey W. McLaughlin, Implementation Realities and Evaluation Design (Stanford, Ca.: Institute for Research on Educational Finance and Governance, Program Report No. 84-B1, 1984).

4. Dorothy Routh, "A Review of the Master Teacher Program, Merit Schools, and the Extended Day" (Tallahassee: Florida Legislature, 1986).

Chapter 3

# REASONABLE EXPECTATIONS FOR EVALUATION

Michael W. Kirst

with contributions by Kimberly Ford and Worku Negash

The current evaluation task facing state policymakers is particularly difficult. The 15 years from 1965 to 1980 were spent perfecting several techniques of educational evaluation, especially program evaluation. Just as evaluators began to understand the limits and strengths of these techniques, however, state governments from 1983 to 1986 changed the types of state interventions, from specific categorical programs to comprehensive education reform packages.

The totality of all the initiatives in many states is not amenable to an evaluation approach used before by any state. Many comprehensive reform packages contain numerous programs, incentives, mandates, and suggestions. A list of the 80 reforms contained in California's 1983 education improvement act for example, includes the following:

Administrators
Notice of Termination
Pilot Projects for Improvement
Seniority Rights
Training and Evaluation
Adult Education
Agricultural Vocational Education
Apprentice Program
Bilingual Exit Criteria
CAP Tests
California Academic Partnership Program
Categorical Funds
Classroom Teacher Instructional Improvement Program
Commission on School Governance and Management
Computer Curriculum
Continuation Schools
Counseling for 10th Graders
Curriculum Standards
Economic Impact Aid in High Schools
Education Improvement Incentive Program
Facilities/Property
Golden State Examination Program
Graduation Requirements
High School Accreditation
Innovative Local Experiments to Strengthen Personnel & Management
Instructional Materials
Instructional Time
Longer School Year
Mentor Teacher Program
Minimum Day
Minimum Teacher's Salary
Opportunity Classes/Programs
Regional Occupational Centers/Programs
Revenue Limits
School Improvement Program
Small School District Funding
Small School District Transportation
Special Education
Specialized High Schools
State Department of Education Guidelines
Students
- Counseling for 10th Graders
- Discipline (Suspension/Expulsion)
- Graduation Requirements
- Minimum Day
- Number of Courses, Seniors
- Promotion/Retention
- Truancy
- Unexcused Absences

Summer School
Teacher Education and Computer (TEC) Centers
Teacher Shortage Loan Assumption
Teacher Trainee
Teachers
- Collective Bargaining
- Credentials
- Dismissal/Suspension
- Evaluation
- Evaluators
- Layoff
- Probationary
- Seniority
- Staff Development
- Test Before Reassignment

Transportation
Urban Impact Aid
Waiver of Education Code for Counties
Work Experience

Source: California State Department of Education, 1984

Another problem that makes evaluation complex and lengthy is the absence of uniform or standardized state policy. State policies vary enormously, from minimal intervention techniques in local policies (New Hampshire) to the extensive state prescription used in Hawaii, which has no local school districts. The concept of "state policy" is further clouded by the implementation variation among local education agencies that have been affected by a particular reform. The appropriate unit of analysis is what takes place in the classroom, the individual school, as well as in the entire school district. Consequently, differential implentation at the local level means that "state policy" will not look the same even within the same state. Implementation is an uncertain process that changes, distorts, or improves the orginal state intent, depending on local circumstances and capacity. No "state policy" spans all the diverse state and local contexts in the United States.

Indeed, evaluation is probably not the correct term to use in the case of the types of omnibus state reforms currently underway. There probably will be no overall judgements concerning whether these reforms are "working" or whether they are "a complete failure." Both these extremes are unlikely complex because the various reforms will be partially implemented and have so many complex interactions that an overall summary judgetment is exceedingly difficult. The term assessment is also inappropriate because it often means reliance on an overall state or local achievement test. These new state reforms require multiple measurements, not a single, overall simplistic indicator like SAT scores. These multiple measurements could include everything from measures of improvement in the quality of teachers to changes in high school course offerings or staff morale. Moreover, statewide indicators should be supplemented with particular local indicators that can probe areas such as school climate and local willingness to implement state activities.

The most difficult and perhaps most unlikely evaluation stage to be reached in the near future is an accurate analysis of the cause and effect of these various education reforms. If a state reform program has between 40 and 80 reforms (Illinois and California, respectively), how can one possibly determine

the cause and effect of so many different interventions coming into the local system? It is comparable to prescribing many different medicines for a patient and then trying to isolate the effect of one pill. Some of these drugs would be taken orally, some would be injected, and all would merge somewhere in the body. Sorting out which medicine did what would be difficult, even for medical people who have better experimental conditions for measuring cause and effect than educators have for measuring the educational system. The human body is complex, but not nearly as complex as sorting out the effects of 80 reforms directed at 7,000 local schools in California.

Kratwohl (1985) uses the following example to demonstrate various level of causality.[1] When a moving car suddenly comes to a halt and the gas gauge reads "empty," one would conclude that because the gas tank is empty (cause), the engine quit (effect). This obvious conclusion seems reasonable and is probably true. However, to answer this question, other alternatives must be considered. Some alternatives, such as a damaged fuel pump, may relate to the presumed cause. Others, such as a loose battery cable, may be completely unrelated to the fuel problem.

Commonly, individuals draw conclusions about cause/effect relationships without even considering whether their problem definition is adequate. Most people's implicit model of problem solving resembles that in Figure 1.

Figure 1.

Problem-Solving Model

PROBLEM DEFINITION SOLUTION IMPLEMENTATION

FEEDBACK

Causal inference occurs at the problem definition stage. No previous attention is given to finding the real problem -- the distinction between a fuel problem and an electrical problem -- so that choosing a solution and implementing it proceed on faulty grounds.[2]

Level of Causality

A second question concerns who is to blame for the effect of running out of gas. For instance, the tank may be out of gas, but the issue of causality remains open. Is the problem carelessness in not checking the gas gauge; a "devil-may-care" attitude about automotive maintenance; or is it due to parents who did't emphasize responsibility for such matters?

Often, we see only the last link in the causal chain, the most immediate or observable element. Many other options exist: the most blameworthy, the most remediable, the least obvious, the most fundamental.[3] Which issue should the evaluator choose? For example, a psychologist, trained in the psychoanalytic tradition, may conclude that the real issue goes back to parental interactions. In an evaluation setting, this choice would lead to vastly different problem definitions, methodological decisions, and ultimately results.

Causal Patterns

The third question, "What is the causal pattern between A and B?," implies yet another source of complexity. To illustrate several simple patterns, Krathwohl used another example.[4] One assumes that a spastic child's slurred speech and awkward movements (A, the cause) cause her to be dependent (B, the effect) on others. This simple relationship can be represented with symbolic notation.

A --> B

However, the relationship could be defined in terms of B causing A.

A <-- B

Using the example, a child's dependency (B) may invoke in adults expectations of helplessness. The child's awkward behavior and speech simply may be a response to those expectations.

Perhaps a more plausible theory is that the dependency and expectations interact and reinforce one another.

A1 --> B1

A2 --> B2

The child's learned helplessness encourages greater efforts by the adults, which in turn reinforces the child's perceived need. The question of causality is even more complex in this recursion, and one is still uncertain about the root cause.

Assessing current education reforms requires new data collection and management techniques and alternate means of conceptualizing interrelationships among the various state interventions. Until these difficult tasks are negotiated, little solid cause/effect research may proceed. It is likely, then, that cause and effect research will not occur until the latter stages of state education reform assessment.

## TOP-DOWN VERSUS BOTTOM-UP EVALUATION: The Need to Use Both

One improvement that evaluators have made in their strategy has been to combine the concept of "top-down evaluation" with that of "backward mapping." The initial evaluations of federal programs in the 1960s were almost exclusively top-down evaluations. They followed a congressional statute through its reformulation into federal regulations and guidelines. These national guidelines then flowed from the federal agency to state agencies to the local central school district, and from there to the local school district and classroom. Differential implementation and discrepancies between federal goals and state-local implementation were carefully documented as state and local educators displaced the federal goals with policies reflecting goals of their

own.

A state statute, administrative order, or school board resolution is rarely permanent. Nor is any state statute self-executing, for how, or even whether, it is executed is a function of numerous intervening forces -- the attitudes and resources of its opponents and supporters, the energy with which political authorities seek compliance, and so on. Thus, a statute arrives in the cold world like an infant, alive with infinite possibilities. In another sense, state law is a memorandum to the future. As Emerson commented:

> The law is only a memorandum. We are superstitious, and esteem the statute somewhat: so much life as it has in the character of living men is its force. The statute stands there to say, Yesterday we agreed so and so, but know feel ye this article today? Our statute is a currency which we stamp with our own portrait: it soon becomes unrecognizable, and in process of time will return to the mint.[5]

State policymakers need to learn what happens after the legislature "makes a law" or a state court issues an order. Each of these is taken as a political decision at one point in time -- "Yesterday we agreed so and so." That decision will differentially influence the values and resources of its supporters and opponents, who, interacting with the initial policy decision, will advance different interpretations of its administration.

The dynamism of this process is made greater by the fact that the United States operates within a federal system. There is no distinct compartmentalization of sovereignties and powers among the three levels of nation, state, and locality but an interlocking pattern of cooperation and conflict. In Grodzin's phrase, it is not a three-layer cake but a marble cake. When a systems framework is imposed upon the mosaic of subsystems constituting federalism, the policy flow can be represented symbolically by Appendix I.

Does a state law, when it finally reaches those it is designed to affect, actually achieve the desired legislative goal, e.g., more qualified teachers of science or language, more students better prepared to start school or express their ideas? During the 1957 cold war phase of American foreign policy, many Americans were obsessed with the notion that the Russians in their foreign

ventures and space developments were "ahead of us." Much of this sense of stress focused upon deficiencies in our supplies of particular technical skills. Our diplomats were seen as lacking language skills in the countries where they represented us, and Soviet "leads" in the "space race" pointed up a purported gap in the number of scientists.

Much of this general distress was sharply focused when the Soviets put a satellite into orbit in 1957. This general stress crystallized into demands to "do something" about this condition. The media and Congress were filled with varied indications of unhappiness that a nation America conceived of as composed of peasants and outhouses had put scientific hardware into space.

Out of these demands, Congress passed the National Defense Education Act of 1958 (NDEA). Funds were provided to encourage students to adopt a teaching career in science, math, or foreign language, and to increase the number of teachers in general, an occupation then considerably under-supplied. Student loans would be "forgiven" if the graduates taught in the selected field or worked in the worst -- or "disadvantaged" -- schools. NDEA used an economic motive to induce social change and several states in 1983-86 have incorporated these ideas into a state loan forgiveness program for new teachers.

As the federal law interacted with the national environment, particularly with that of the school system, what were its effects? A carefully drawn study in 1971 found that the "forgiveness" feature of the law had had little effect upon student career choices, nor did it move prospective teachers in larger numbers into "disadvantaged" schools.[6] The reason was that few were motivated in such decisions by money alone.

One stark indicator is that, despite the $1.3 billion authorized for loans, only $94.8 million were spent for the "forgiveness" feature, and only $30 million of that will be cancelled in future years by graduates doing what the act was primarily designed for. In short, as Russell concluded, "This nation's initial major attempt at manpower planning has not been an effective vehicle to

affect significantly the quantity, quality, distribution, or retention of prospective teachers."[7]

The concept of backward mapping, on the other hand begins the whole process not from the state or federal level, but from the classroom. In backward mapping, one would look at the problems and conditions in the classroom, understand those well, and then build the evaluation back up through the delivery chain to the school site and eventually to the state capitol if that is where the intervention began. Backward mapping tends to favor local flexibility, whereby local school officials, probing their own problems and their own conditions can devise programs around those, rather than fitting state and federal regulations into an unmalleable local context. The most effective evaluation technique merges both top-down and backward mapping viewpoints in the same analysis and field study.

A number of well-intentioned state reforms also may have the unintended consequence of depressing teacher morale substantially. This morale decline should lead state policymakers to question their basic reform assumptions and search for alternatives to their initial approaches. The concept of unanticipated and unknown consequences is particularly relevant, given the experimental nature of many state reforms such as career ladders.

Throughout, this report will stress the idea of rethinking basic policy assumptions and intervention strategies as an outcome of evaluation. For example, Sykes has advanced a basic two-fold, alternative strategy for attracting and retaining high-quality teachers.[8] He contrasts the approach of "screen", such as teacher competency testing and raising certification requirements, with "magnets" to attract better teachers through, for example, salaries, working conditions, and sabbaticals. State policymakers may find through their evaluations that they are relying more on one of these strategies than on the others. In the early 1980s, it appeared that states were oriented almost exclusively toward raising a screen for entry into the profession. Lately, they have turned more toward magnets, including raising minimum salaries in many states. But what is the proper mix between the alternatives of screens

versus magnets? Moreover, given limited resources in state money, what is the most effective mix of these strategies? In chapter 6, we propose a specific cost/effectiveness technique to deal with this question.

One of this reports's themes is not to expect too much from evaluation, but another is that evaluation can be exceedingly usefully in short and long-range planning. In the mid 1970s, state legislatures tended to have unrealistic expectations about the "final answers" that evaluation could provide. California, for example, began a new program in 1975 using local school site councils, called the School Improvement Program (SIP). This program created school site councils of parents, teachers, and administrators to provide bottom-up programs that the schools thought were the most useful. Each school council was funded through a competitive grant of about $100 per pupil to use flexibly as it perceived the greatest need. State legislators wanted to know what the pupil achievement impact of SIP was on the standard California statewide test. Legislators thought that this single SIP intervention in a complex system could be measured by looking at the overall level of elementary achievement achievement scores and changes in those scores from year to year. In fact, statewide achievement is affected in California by numerous forces. California had 20 categorical programs of its own in the mid-1970s, as well as an additional 15 to 20 from the federal programs. Moreover, California's pupil population had become more minority oriented, with a particular growth in limited-English-speaking students.

Yet, the California Legislature wanted to know whether The School Improvement Program, a minor intervention in terms of the overall education context, was the cause of changes in the overall test scores for a sample of California third graders. Such a question seemed reasonable at first glance, but a more complex understanding of the many influences coming into the school districts, as well as the varied local implementation settings, suggests that this test measure is inappropriate. But when the legislature shifted its evaluation dollars to the factors and policies that were helping or impeding local school site councils, that approach began to pay off. Several state mandates were eliminated in order to provide the school site councils with more

flexibility.

## COMPARATIVE STATE EVALUATION

Comparative state evaluation across different state approaches is a concept that will be useful in this endeavor. By comparing and contrasting implementation, process, and outcome effects among the different states, more can be learned than just by studying one state in depth. Since state reforms vary in content and objectives and the state and local context varies as well, policymakers should probe for the differential effects of the same reform in different state settings. Alternatively, states can examine the differential impact of different state reforms on the same types of local school districts in many states (for example, small rural districts or inner-city school units). This suggests that comparative evaluation be funded by agencies that can put forth more effort than a single state. State collaboration through such organizations as NCSL or the Education Commission of the States is an obvious advantage. Moreover, networks of state evaluators must be built so that they can systematically compare and contrast their evaluation approaches and results. A large role is also implied for the federal government, whose effort can span numerous states. By comparing and contrasting state approaches, one can undoubtedly learn a lot more than through just looking at any one state approach. In effect, states are providing a large-scale, unprecedented "natural laboratory" for evaluators who want to understand the implementation processes and outcomes of large-scale reform. This natural variation across states and local districts can be exploited to provide substantial insights on policy.

## IMPLEMENTATION AS A FOCUS OF EVALUATION

McLaughlin defines implementation as "the process whereby programs or policies are carried out; it denotes the translation of plans into practice."[9] As recently as the late 1960s, implementation analysis was often ignored. For example, a study by the G.E. Tempo Division of Title I of ESEA assumed that local districts followed the strict letter of the federal regulations in implementating the program. The evaluators assumed that the target group pupils

were identified and served according to the federal guidelines. Consequently, they spent no time looking at the so-called "black box" between inputs an outputs. This black box embodies the implementation process. The evaluators discovered no measurable change in pupil achievement scores and declared the Title I program a failure. Subsequent evaluators, however, looked at the implementation process and found that the intended beneficiaries (low-income children) were not receiving the services provided by substantial amounts of the federal money. Therefore, the first necessity was to straighten out the implementation process and only then, evaluate its incremental improvement. Only then could the evaluators explore the changes in pupil outcomes, which should include much more than test scores. In the South, Title I, for example, provided huge amounts of health, clothing, and nutrition programs; therefore, the initial outcome measures should have been on the health of the pupils rather than on their reading scores.

This type of experience has led evaluators such as McLaughlin to the following conclusions:[10]

1) Treatment effects from state reform often are indirect. For example, an attempt to implement a model state curriculum will not be translated by local school districts directly into their curricular guides, or by teachers into their detailed lesson plans in a straightforward, uniform fashion based on the recommended state curricular content. The curriculum guide will influence some teacher thinking and change some of the content they present.

2) Implementation choices made by local school systems and teachers often will dominate and overwhelm intended educational reform outcomes of state policymakers. Education has always been a "bottom-heavy" enterprise, with the classroom teacher exercising a "pocket veto" once the classroom door is closed. This realization is particularly crucial in cases where states are trying to increase academic standards as a prime educational outcome.

3) Implementation is a multistage development process that must be explored over a fairly long period of time. Initial implementation attempts may

not have high fidelity to state reform goals, but continued state pressure may bring the implementation closer to the intended reform goals.

4) Local school implementers pursue multiple and often competing goals. State reform efforts must compete with local pressures within the local school system, and consequently, will be bent and deflected according to those distinctive local goals. A prime local goal, for example, may be to maintain a robust and rounded vocational education program. But state policymakers, in pressing for academic classes, are cutting back on vocational programs. The outcome at the local level from these multiple and competing goals may be some compromise between instigating state reforms and supporting local vocational education.

5) Decisions made closest to the child are the most influential. This approach to evaluation suggests the importance of backward mapping by focusing evaluation resources on the point where the state reform encounters the child, i.e., the classroom, rather than relying totally on top-down assessment or the precision of state regulations.

All these conclusions imply that state evaluators cannot "prove" that a reform has worked or has had a particular impact in numerous schools. The different reforms will behave and combine differently in various local settings.

This implementation viewpoint suggests that the most appropriate unit for states to evaluate is the individual school as a whole, not just a particular program or intervention such as teacher testing or a new textbook program. At the end of this report an intensive analysis is called for over an extended period of time (five to 10 years) of a sample of local schools in a state. Only through such in-depth longitudinal efforts can the effects of this wave of state education reform be accurately assessed. The goal of state policymakers should be to learn more about reform for future reforms. The omnibus reforms need to be shorten, shifted, and amended. Consequently, state policymakers should use evaluation as an incremental learning process whereby they gradually know more what is happening at the local level as a result of omnibus bills.

AN EXAMPLE:

STUDENT DROPOUT RATES

The dropout issue is emerging as a prime consideration and concern in many state legislatures. But this issue is a classic example of a number of the evaluation problems previously mentioned in this report. For example, no state has an accurate, valid measure counted as a dropout. Dropout statistics encompass pupil attrition and mobility between school districts rather than leaving the entire school system. Dropouts do not send letters to the school system saying, "I hereby drop out." The outcome measure of a dropout is, therefore, suspect from the beginning. But even if one can settle the outcome measurement issue or improve it drastically, the interaction effect raised earlier still exists. One can assume an increase in the number of dropouts in a particular state. The press and critics of educational reform may blame this on the educational reform program. Some of the allegations may be that higher standards and fewer course choices lead to more dropouts. But what is the input-output or cause-effect relationship between the state's reform and the number of dropouts? This is no easy question. First, one cannot be sure local school districts have even implemented state reforms with fidelity to the original objectives. If the local school districts have not implemented the reforms, then they cannot be the cause of students dropping out.

If policymakers discover that local school districts have implemented reforms with rather high fidelity to state objectives, then the multiple cause and program interaction issues arises in a major way. For example, an increase in the number of dropouts could be caused more by changing pupil populations rather than anything to do with state reforms. Traditionally, Hispanic pupils have a higher dropout percentage than non-Hispanics. A large increase in disadvantage pupils as a percentage of 17-year-olds could be causing students to drop out in large part, with state reform playing little or no role. On the other hand, the number of dropouts could be increased by a particularly robust economy, whereby 16-year-olds realize that they can get good jobs immediately; therefore, they feel no need to graduate. Again, reforms could be fingered as

the culprit in this distressing outcome whether they are the real cause. Finally, it is possible that reforms will motivate of push out a number of kids. Although understanding these possibilities is not easy, evaluators do know how to "hold other things constant" to some extent; thus, they might begin isolating the effects of educational reforms from economic and demographic changes, for example.

## STANDARDS FOR JUDGING WHETHER AN EVALUATION IS SOUND

Several authors have provided standards by which to judge the quality of evaluations, but these standards frequently are based on single programmatic interventions or a much more limited set of state policies than is embodied in the current state and local education reforms. Some respected canons of educational evaluation standards, however, will be useful to keep in mind even though they might not be fully applicable to the current state reform effort.[11] State policymakers should seek evaluations that have distinguishing characteristics such as the following:

1) Other observers can duplicate an educational program by using the same evaluation techniques, instruments, and measurement devices.

2) The evidence on the results of an educational program or reform is subject to tests regarding whether the finding could have occurred without state intervention.

3) The state policymaker has information on whether funds for the reform were used efficiently.

Evaluations always must consider the range of competing and alternative consequences of various state policies.[12] Such standards are difficult to attain even when the state reform is a direct intervention such as a school breakfast program on the nutrition of children. The standards are more difficult to meet when one considers an omnibus state strategy with dozens and dozens of reforms interacting within a single school site.

## TAILORING EVALUATION TO SPECIFIC CONSUMERS

Educational evaluators have learned a great deal in the past 20 years about how to tailor their presentations and analyses to the particular needs and time constraints of policymakers. Legislators need to seek out "information brokers" who can reduce the technical evaluation reports to easily understandable language. Moreover, these brokers take into account not only the realities of legislative decision making, such as short timelines, but also the policy needs of different legislators. Some public officials, for example, are much more receptive to numerical analysis than others, and legislators differ greatly in how they process information. Some focus more on details and are more inductive and interpretative. Others desire only oral presentations that provide general conclusions. A crucial part of evaluation strategy in the legislative setting is matching the evaluation information with a broker who understands a legislator's learning style and motivations.

Good evaluations try to incorporate many different techniques into their repertoires. They may use implementation approaches, cost-effectiveness approaches, and traditional cause and effect studies. Legislators should search for analysts who are flexible in the types of evaluation approaches they use. Evaluators to avoid are the those who have a set technique that they try to impose on every type of evaluation problem. For example, in Chapter 6, cost-effectiveness analysis is recommended, although this type of analysis is not adaptable to situations where several programs have significantly different objectives.

A related issue arises when evaluators disagree. Everyone is familiar with the situation where two people review a novel or movie and come to very different assessments. Legislators frequently are faced with "partisan evaluation." Two competing interest groups, for example, will produce evaluations showing diametrically opposed results from a legislative intervention. Both will have hired an outside evaluator who is using accepted evaluation research techniques. What does a lawmaker do in such circumstances? First, he or she must examine the theoretical and philosophical motivations from

which the evaluator began the task. An evaluator working for the National Education Association may have a different philosophy, orientation, and approach than one employed by the National School Boards Association. If evaluators disagree, a legislator could turn to a panel of relatively uncommitted third-party judges. This technique works best, however, when a legislator also knows the philosophy and research orientation of the various panel members. One way to do this is to deliberately pick a diverse panel.

In the end, a legislator is similar to a connoisseur of art or literature. A legislator must make judgments, not just impose his or her preferences, and these judgments must be based on the data produced. But it is still likely that different legislators will make different judgments, each seemingly equally based on the same data. Critics reviewing the same art or movies end up disagreeing even though they start from similar preferences. The best advice is that a legislator should obtain several different perspectives, but realize that evaluation is a process of learning gradually about reform in order to improve it -- what evaluators call "formative evaluation." Formative evaluation is a preliminary assessment concerning how the state reform is progressing rather than a final summative judgment on its overall worth. Formative evaluation enables legislators to alter their policies during the developmental phase of reform rather than wait for five to 20 years to elapse before any kind of summative statements can be attempted. Summative statements in this round of educational reform are likely to be very difficult to make.

While evaluation can assist in making some discrete decisions about the future direction of state reform, it also has several other uses. Evaluation can be used in conflict resolution. Evaluation signals that a state reform program is under study and subject to further negotiation and compromise among various interest groups involved in formulating the state education reforms. When interest groups with conflicting views agree to evaluation as a partial test of a state reform's worth, they commit themselves to some compromise solution related to the evaluation results. Moreover, the very act of participating in an evaluation may spur consideration of new approaches by frontline practitioners such as school administrators and teachers. Just

participating in an evaluative effort in a school system can help local schools officials and state policymakers clarify their standard operating procedures, goals, and policies, and revise those that are ineffective. Indeed, even discussing how to measure a state reform program's success is a beneficial exercise in goal formation.

In sum, a specific evaluation rarely provides a definitive answer of whether a state education program should be cut, drastically increased, or eliminated. Often the impact of evaluation on state policymakers is more indirect and subtle. A series of related studies, concepts, and legislative hearing permeate the policymaking process. Evaluation percolates through the legislative process and shapes the way in which legislators think about educational issues. Consequently, evaluation provides enlightenment and perspective in addition to some specific guidance about specific short-term decisions.

This view of evaluation is summed up well in the following propositions:

1. Evaluation is a process by which society learns about itself.

2. Evaluations should contribute to enlightened discussion of alternative plans for social action.

3. Evaluation is a handmaiden to gradualism, both conservative and committed to change.

4. An evaluation of a particular program is only an episode in the continuing evolution of thought about a problem area.

5. A theory of evaluation must be as much a theory of political interaction as it is a theory of how to determine facts.

6. The hope that an evaluation will provide unequivocal answers, convincing enough to extinguish controversy about the merits of a social program, is certain to be disappointed.

7. Goals are a necessary part of political rhetoric, but all social programs, even supposedly targeted ones, have broad aims.

8. Legislators who are quite sophisticated about their own reasons for keeping goal statements lofty and nebulous unblushingly ask program evaluators to be specific.

9. Unfortunately, whatever the evaluator decides to measure tends to become a primary goal of program operators.

10. Much that is written on evaluation recommends some one "scientifically rigorous" plan. Evaluations should, however, take many forms, and less rigorous approaches have value in many circumstances.

11. Results of a program evaluation are so dependent on the setting that replication is only a figure of speech; the evaluator is essentially an historian.

12. A strictly representative sample may provide less information than a sample that overrepresents exceptional cases and deliberately varies realizations.

13. In project-by project evaluation, each study analyzes a spoonful dipped from a sea of uncertainties.

14. The evaluator should allocate investigative resources by considering four criteria simultaneously: prior uncertainty about a question, costs of information, anticipated information yield, and leverage of the information on subsequent thinking and action.

15. The evaluator is an educator; his success is to be judged by what others learn.

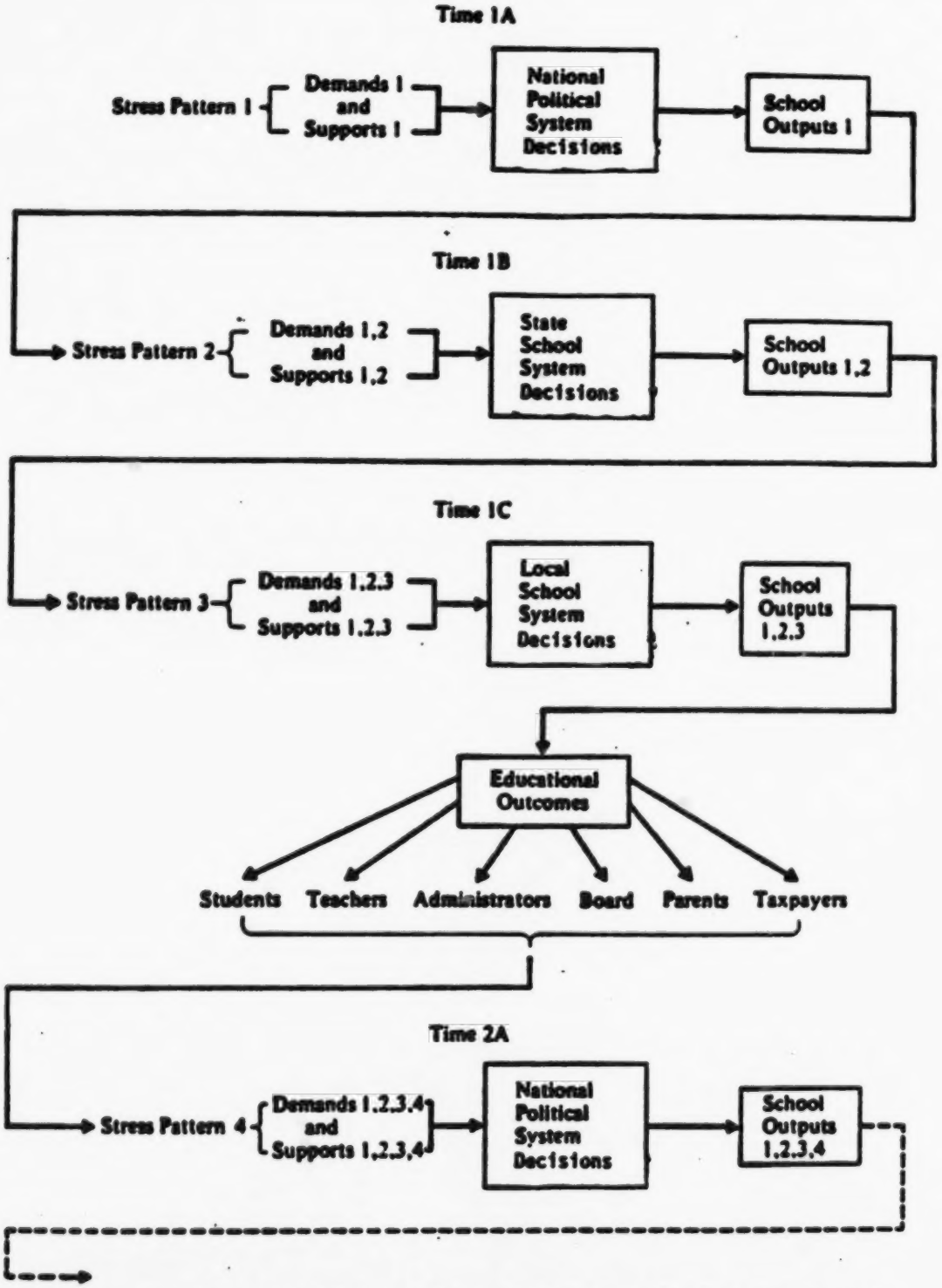
---

We endorse the above propositions from the book by Cronbach and his associates as a good summary of this chapter.

From Lee Cronback et al., Toward Reform of Program Evaluation (San Francisco: Jossey Bass, 1981).

APPENDIX I

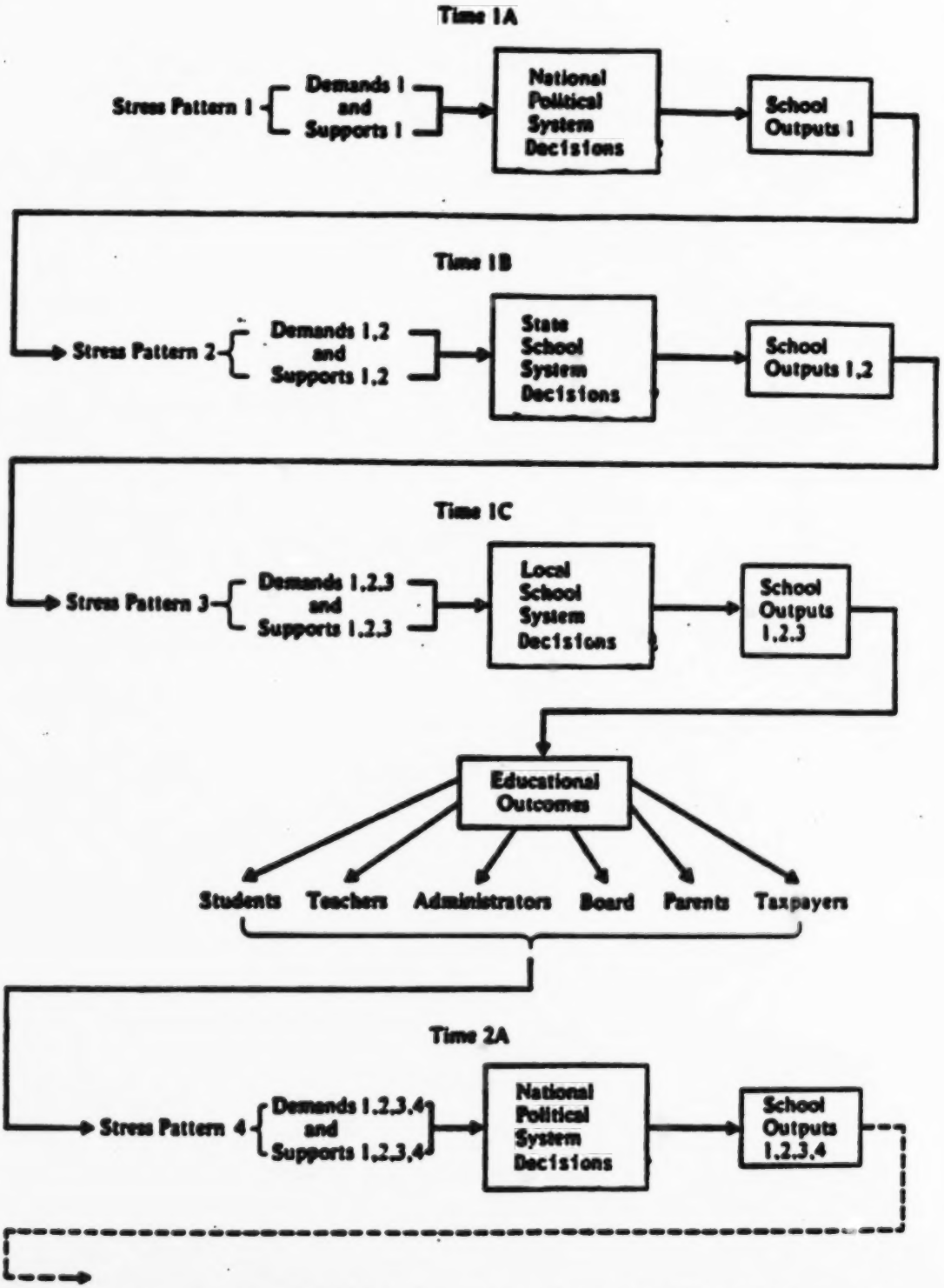
The Flow of Policy in a Federal Education System




Source: From Frederick Wirt and Michael Kirst, Political and Social Foundation of Education, Berkeley:McCutchan, 1975, p. 149.

APPENDIX I

The Flow of Policy in a Federal Education System




Source: From Frederick Wirt and Michael Kirst, Political and Social Foundation of Education, Berkeley:McCutchan, 1975, p. 149.

## Chapter 3

## FOOTNOTES

1. David R. Krathwohl, Social and Behavioral Science Research (San Franciso: Jossey-Bass, 1985), p. 212.

2. Harold J. Leavitt, "Beyond the Analytic Manager, Parts I and II," California Management Review 17, 3-4 (Spring and Summer 1975): 5-12 and 11-21.

3. Krathwohl, 1985, p. 212

4. Krathwohl, 1985, p. 215.

5. As quoted in Eugene Eidenberg and Roy D. Morey, An Act of Congress (New York: Norton, 1986), preface.

6. William J. Russell, "Occupational Choice: An Analysis of the National Defense Student Loan Program," Ph.D. dissertation, University of California, Berkeley, 1971.

7. Ibid., p. 68.

8. Gary Sykes, "Contradictions, Ironies, and Promises Unfulfilled: A Contemporary Account of the Status of Teaching," Phi Delta Kappan, October 1983, pp. 87-93.

9. Milbrey W. McLaughlin, Implementation Realities and Evaluation Design (Stanford, California: Institute for Research on Educational Finance and Governance, program Report No. 84-B1-1984), p.3.

10. Ibid., pp. 6-13.

11. Adapted from P.H. Rossi, Howard E. Freeman, and Sonia R. Wright, Evaluation: A Systematic Approach (Beverly Hills: Sage, 1979).

12. Ibid., p. 50.

Chapter 4

# THE USE OF TESTS IN MEASURING THE EFFECTIVENESS OF EDUCATIONAL PROGRAMS, METHODS, AND INSTRUCTIONAL MATERIALS*

Ralph W. Tyler

## Editor's Introduction

Legislators often use test scores to make judgements about whether a state initiative is working and helping children learn more. Tests seem simple but actually are complex, confusing, and often misinterpreted. Different judgements and see different trends from the same test scores. Each year a typical student takes many kinds of tests: Some are mandated by the state, others required by school districts; some exams are prepared by a teacher, others are special exercises recommended by a specialist.

Most achievement tests in the United States are based on norm-reference concepts that encompass the familiar bell-shaped curve. In other words, the achievement test is designed to sort out the differential abilities of various pupils across a distribution. This approach was a particular problem in measuring the impact of Title I of the Elementary and Secondary Education Act (ESEA) in the early 1970s. If most pupils began to know items on a test, then the test makers would eliminate those items because they would not distribute the students over the bell-shaped curve. Consequently, programs such as Title I ended up like a dog chasing its tail. As soon as most pupils knew test items, they were deleted and more difficult questions were inserted in their place. The low achievers could not show any increase in performance over time. Consequently, this report explores the arcane world of educational testing. The objective is to alert state policymakers to the complexities of testing and make them particularly concerned about overly relying on these measures.

Nevertheless, testing programs can provide several useful types of information. For example, the National Assessment of Educational Progress (NAEP), administered by the Educational Testing Service for the U.S. Department of Education, repeatedly has stressed that pupils do not have an adequate grasp of the higher order and create thinking skills in areas such as reading and mathematics. This conclusion can be drawn from test and from an analysis of testing trends; however, it is unlikely that a statewide

---

* Reprinted by permission of the publisher, McCutchan Publishing Corporation Berkeley, CA, from Crucial Issues in Testing. by Ralph W. Tyler and Richard Wolf, editors, Copyright 1974.

test can measure the effect of a single reform such as a new statewide curriculum, when so many other factors are changing simultaneously within the state and local education system.

Ralph Tyler is one of the most experienced and eminent authorities on testing. He focuses here on the uses, limits, and misuses of testing for reaching policy conclusions about new programs and initiatives. For example, often existing tests are employed that do not properly reflect the different educational objectives of a new program. The test items are designed for the "old approach." Moreover, frequently, there is not a close match between the new curriculum and the test content based on the old curriculum. Some tests include low level basic skills such as comprehension of short paragraphs while others require the ability to understand the author's tone and mood. In the pages that follow, Tyler explores these and other differences in tests and cautions against simplistic policy assumptions based on tests.

Since 1925 the accepted design for studying educational effectiveness has been, at some initial point, to give an achievement test called a pretest, and, at a final or later point, a comparable post test. The gain in the mean score was taken as a measure of the educational effectiveness of the program, method, or instructional material. In case a pretest could not be given, the final scores could be compared with test norms or with the scores obtained from comparable groups not following the program method or material under study. Recently, however, the use of typical standard tests for this measurement has been seriously questioned on several counts.

Achievement tests were designed to appraise individuals in terms of their deviations from the mean of the population to which they belong. The score usually reported for an achievement test is an abstract number that has no concrete reference. It is usually derived from a number which is the sum of the number of exercises answered correctly. Thus it is two steps removed from reporting what tasks the individual has performed correctly. If typical

achievement tests are employed to appraise the educational accomplishments of a school or school system by using the mean or average score of the students, this is yet a third step removed from the actual performance of individual students on particular test exercises. Comparisons can be made with the mean score of the norming population of students and with mean scores of other groups, but these calculations do not answer questions about what students have learned. Furthermore, the method of constructing the test to maximize items that reveal individual differences among students makes it impossible to answer questions about what all or nearly all students have learned and what important things have been learned by only a small percentage. The exercises on such tests are not selected to form a representative sample of what student have learned. Instead, they are a sample of items that clearly differentiate among students. Since the purpose of assessing schools or school systems is to ascertain what various populations of students are learning, so-called criterion-referenced tests are being constructed to serve this purpose more directly than the norm-referenced tests.*

The effort to appraise institutions rather than individuals involves problems of sampling populations and subpopulations of students enrolled in the institutions. To make an efficient assessment of the learning of students within an institution does not require that every student be tested if a representative and reliable sample of students can be selected for testing. This procedure raises some technical issues that are currently being studied. The sampling problems are different from those commonly considered by test specialists. Criterion-referenced tests require representative and adequate samples of each of the major things that students are expected to learn, and efficient population sampling procedure requires representative and adequate samples of each population or subpopulation of students in the institution. The first requirement results in many more exercises than are commonly included in a norm-referenced test because the latter test can obtain a representative and adequate sample of items that differentiate among students using very few items from each of the divisions of things to be learned. The requirement that a representative and adequate sample of each population or subpopulation of

students enrolled in the institution be utilized usually results in a great reduction in the number of students responding to each exercise in comparison with the practice in norm-referenced testing, in which everyone is usually included because of the use of these tests to compare individual students.

---

*The test used in the National Assessment of Educational Progress (NAEP) is a good example of criterion-referenced tests.

**Criticisms of Standard Achievement Tests**

One criticism is that these tests do not properly reflect the educational objectives of the new program, method, or instructional material. Usually, new educational programs are constructed to gain certain new educational objectives, to alter the emphasis on earlier aims, or, perhaps, to achieve both goals. For example, the "new mathematics" programs devised in the 1960's included as a major objective the development on the part of the students of an understanding of and an ability to use ways of thinking characteristic of mathematical study. Although the new programs also aimed at some of the objectives of older programs, such as computational skill and the ability to solve everyday quantitative problems, they were not given the same emphasis. Hence, when standard tests in arithmetic were employed to measure the effectiveness of the new mathematics programs, they were sharply criticized because the tests did not include exercises that sampled the students' understanding of and ability to use ways of thinking characteristic of mathematical study.

New methods of teaching and new instructional materials, like new programs, are also commonly developed as a means of attaining new objectives or of giving new emphasis to older ones. Some programs are constructed and some methods and materials are devised to help solve a learning problem that has been identified as critical. For example, after World War II, oral-aural methods of teaching a foreign language were designed to support the new emphasis on oral-aural skills. Just before World War II, the recommended curricula primarily emphasized the development of reading skills. The design of behavioral

modification methods (operand) conditioning and the accompanying materials in primary school reading were a response to the recognition of the critical problems encounted in teaching reading to so-called "disadvantaged" children.

The use of the common published achievement tests to measure the effectiveness of new teaching methods or of new materials that are focused on new objectives is criticized on the same ground as their use for new educational programs, that is, the tests do not properly reflect the educational aims being sought. Another criticism is made when these tests are used to appraise the effectiveness of programs, methods, or instructional materials designed to solve a critical learning problem. The charge is often made that the tests do not furnish a reliable sample of exercises appropriate for the students who encounter such a problem. For example, new programs, new teaching methods, and new materials are often designed to improve the learning of disadvantaged children. Typically, these children score in the lowest quarter of the distribution of scores on a standard achievement test. But the exercises found in most of these tests are clustered around the 50 percent level of difficulty, with fewer than 10 percent of the exercises below the 25 percent level. Usually less than 5 percent of the exercises in a typical standard test are at the difficulty level where most disadvantaged children would make correct responses. This is such a small sample of items that it cannot furnish a reliable estimate of changes in the performance of disadvantaged children unless the new program, method, or instructional materials help to effect an enormous improvement in these children's responses. Standard tests were not designed to provide reliable measures of change at the extremes of the distribution.

A third criticism is leveled when typical achievement tests are used to appraise new instructional materials. The claim is sometimes made that such tests have been largely built by writing items that use only the words or the modalities found in existing materials, and they only provide information about how well students have learned what is presented in older materials. If proper appraisals are to be made, the tests should enable students to show that they have learned important concepts, principles skills, and so forth, independent of their embodiment in the instructional materials.

## Achievement Test Construction Not Consistent with School Efforts

In order to examine these criticisms impartially, it is necessary to review the bases on which achievement tests are constructed. School achievement is a term commonly used to mean that students are learning what the school seeks to teach. Learning is usually defined today as acquiring new patterns of behavior through experience. The efforts of the school are focused on developing certain patterns of behavior that are considered important to help students participate constructively in society and realize more fully their own personal potential. The school curriculum is designed as a set of experiences that are expected to stimulate students to attempt these patterns, to guide their efforts, and to continue the learning activities until the desired patterns of behavior have become established. The purpose of achievement testing is to ascertain whether, in fact, the students have acquired the desired behavior. A test presents situations intended to evoke the desired behavior and then affords an opportunity to record the performance. Since there are many situations which a desired pattern of behavior can be expressed, a test can present only a sampling. The selection of situations (test exercises) is critical in constructing an achievement test since a biased or inappropriate sample will not furnish dependable evidence as to the adequacy of the student's behavior. The procedure commonly followed is an important source of criticism when available achievement tests have been used to measure the effectiveness of an educational program, a method of teaching, or a set of instructional materials.

The most frequent criticism, which was mentioned earlier, is that a given achievement test does not reflect the particular objectives of the educational program, method, or instructional materials. This is, the patterns of behavior that the test attempts to sample are not precisely the same as the patterns which the program, method, or materials are designed to help students acquire. For example, a program in language arts may be designed to help students write more fluently, with less restriction in content, while the achievement test used might be a sample of exercises involving primarily spelling, grammar, and punctuation. Or, as another example, the teaching method to be appraised might be planned to help students conduct inquiry in science, while the achievement

test used might contain a sample of exercises primarily requiring recall of information about science. Tests designed for the national market are constructed to sample topics common to the curriculum of most school systems. For this reason they may not include exercises that reflect those behavior patterns that are the aims of new programs, methods, and materials. A less commonly noted variant of this problem occurs in the case of behavior patterns that require several years for adequate development. School programs differ in the learning sequences followed to help students acquire such patterns. For example, there are several different sequences used to develop reading comprehension in the primary grades. Some schools begin with a "look-say" method using a carefully controlled vocabulary to develop competence in word recognition before students are trained in phonics and other techniques for attacking new words. Other schools begin with a phonic or linguistic approach. A test designed for use in the first grade that properly reflects the curriculum sequence of a "look-say" approach would be inappropriate as a measure of what pupils had learned who were receiving initial training in phonics or vice versa. A test designed for use at the end of the primary grades is more widely appropriate, however, because most reading programs are planned, by the end of the third grade, to help students comprehend a variety of materials without highly restricted vocabularies. As another example, some arithmetic programs begin with counting and then addition, followed later by subtraction; others treat addition and subtraction together. In the early stages of developing number concepts and computational skills, the particular patterns of behavior practices differ in different programs so that an achievement test used to measure the effectiveness of a program in helping students acquire this behavior needs to be carefully examined to see whether it really furnishes an adequate sample of exercises evoking the appropriate behavior.

The second criticism is derived from certain assumptions made in selecting test exercises. Probably because achievement tests were a by-product of testing for selection in World War I and because they were strongly influenced by assumptions made in intelligence testing, achievement test theory assumed that the outcome of school learning was the gradual development of abilities, and, at

any one grade, these abilities were normally distributed among children. Actually, school programs are based on contrary assumptions. The school curriculum provides for a particular pattern of behavior to be developed at a particular time, and a strong effort is made to help all or most children develop this behavior at that time. For example, addition of whole numbers is presented and practiced at a specified period and not taken up every month or year. The concept of photosynthesis is also dealt with at a particular time in the science curriculum. Furthermore, each teacher tries to help all his pupils master the addition of whole numbers, and each teacher of biology tries help all his students understand the concept of photosynthesis. In general, knowledge that a child acquires piecemeal over a long period of time is often learned incidentally. Schools typically do not focus major attention on such learnings. Yet this kind of knowledge is generally found to be normally distributed among children at a given age. For example, an item of general information that can be obtained from such varied sources as T.V., newspapers, magazines, and conversations is often found to be normally distributed among children of the same age, and the proportion of the age group that acquires this particular skills or facts systematically taught to a given age group. In such cases the distribution is highly skewed for that age, and there is not a gradual increase in the proportion of the group that has acquired this behavior.

In the construction of a norm-referenced achievement tests, such assumptions influence the selection of exercises that approximate the conditions, based on an initial tryout. Exercises are retained when the initial results show that children are normally distributed and when there is an increase in the proportion of correct responses from grade to grade. Furthermore, exercises used in the tryout are eliminated when the distribution of correct responses is highly skewed or when there is not a significant increase in proportion from grade to grade. This procedure makes it likely that the selected exercises are not those on which instruction has been focused but, more probably, items that children learn incidentally. The exercises eliminated after the tryout are more likely to be those toward which major instructional effort has been directed.

The use of these assumptions in constructing norm-referenced tests leads to a common misinterpretation of the norm "grade level." What laymen think it means when a report states that a student is at the fourth-grade level on an achievement test is that the student has acquired some major fraction of the behavior patterns that the school is seeking to teach in the fourth grade. But it does not mean this at all. It means that his test score was the same as the median or mean score of all fourth-grade children in the norm group. In many cases the items answered correctly were not typical of the behavior emphasized in the fourth-grade curriculum of the school in which the pupil was enrolled. Norm-referenced tests are not composed of reliable samples of the things that children are being helped to learn in a given grade but, rather, samples of exercises on which children of a given grade differ markedly in performance. The things that most children are learning in that grade are likely not to be included in the test sample because of the item selection procedures.

The third criticism was that typical achievement tests include too small of a sample of exercises appropriate for appraising the learning of children who markedly deviate from the average. Many new educational programs have been developed to improve the learning of pupils whose learning has been inadequate. For example, hundreds of new programs have been designed to help disadvantaged children whose test scores are far below the mean of the distribution. If the tests are to measure improvements in the learning of disadvantaged children, the items on which they would be expected to improve would be the "easier" ones -- that is, the ones which 75 percent or more of the norm group answer correctly. But the method of item selection maximized the number of items which about 50 percent of the norm group answered correctly because those items provide the greatest discrimination in placing pupils on a continuum from highest to lowest score. On most standard tests there are less than 5 percent "easy items," or those answered by 75 percent or more of the norm group. There are so few items at the level where disadvantaged children can be expected to be learning that changes in their test scores due to improved learning cannot be distinguished from those due to chance variations in performance on so few items. If the effectiveness of an educational program, teaching method, or a set of

instructional materials designed to aid the learning of a special group like disadvantaged children is to be measured, tests are required that include a reliable sample of the behavior that the pupils are being helped to learn. At present, this usually necessitates the construction of a new tests that are properly focused on the specific pattern of behavior.

A fourth criticism is related to the third. A teaching method or a set of instructional materials is commonly designed to improve learning significantly, but not spectacularly. A real improvement of 5 percent to 10 percent in the learning of students is often desired. To measure such differences in the learning of children would require a precision of measurement that is not usually attained or even attempted survey tests. Since survey tests (most standard tests) are usually administered no oftener than once a year and since most schools believe that the time they devote to survey testing should be kept at a minimum, the tests are designed to furnish scores sufficiently reliable to identify learning increments taking place in a year but not those that might occur in a fraction of a year. Hence, most survey tests cannot provide information about the differential effectiveness of teaching methods or instructional materials unless the differences are great.

## Measuring Differential Effectiveness

In order to produce a test that can be used to measure the effectiveness of a teaching method or a set of instructional materials, it is necessary to obtain exercises that sample adequately the several patterns of behavior that students are being help to learn, and, if necessary, over sample the behavior that is thought to be uniquely aided by the method or the materials. If the method or materials are to be compared with others, the test should sample in similar fashion any additional patterns of behavior that the others are aimed to develop so that information can be obtained on the relative effectiveness of the different methods or materials in developing each of the several patterns of behavior that are the aims of one or more of the methods or materials. For example, one method of teaching American history in the 11th grade is to assign small teams of students to investigate features of life characteristic of their

grandparents' generation, each team covering one area of life such as the home, economic activity, political life, play, and leisure, securing data from interviews with senior citizens as well as from printed materials. Each team is expected to prepare a written report and to present an oral one for discussion and criticism of the class as whole. The effectiveness of this method is to be compared with that of textbook reading and discussion.

The test developed for this comparison sampled behavior 1) involving the use of historian concepts in explaining the differences between their grandparents' time and the present; 2) recalling information gained in study about life in the two periods; 3) writing clear and well-organized expositions of a feature of life at the turn of the century; 4) making a bibliography to guide those who might want to investigate this period. Note that the first two sections of the test involved patterns of behavior emphasized by the teacher using textbook reading and discussion method while the third and fourth sections represented additional objectives considered important by the teacher using the team-investigation method. In each section the test included an adequate sampling of the particular behaviors that students were expected to develop in this part of the course activities and, thus, it furnished a basis for the precision of measurement required to appraise relatively small increments of learning.

**Measuring Changes in Behavior**

There is another problems involved in assessing the effectiveness of an educational program, method, or set of instructional materials -- that of measuring changes in behavior. At the beginning of a program many students have already acquired some of the desired patterns of behavior or some parts of the patterns. For them, the question is, did they learn additional desired patterns, or did they complete behavior patterns already begun? Did the educational means being tested aid the student in going beyond the point where he began? The simplest way to answer such a question would seem to require a measure of the student's relevant behavior patterns at or near the beginning of his participation in the program or the method or in his use of materials, and

again at the end. To measure change requires two or more measures at different points in time.

In sum, achievement testing is much more complex than it appears initially. Moreover, it is only one part of a data system that is needed to assess state education initiatives.

## Chapter 5

## NEW DIRECTIONS FOR STATE EDUCATION DATA SYSTEMS

Michael W. Kirst

### OVERVIEW OF THE ISSUES

Evaluation must rely in part on high-quality state data bases. State education data systems appear to be driven in large part by policy trends and priorities. When categorical programs are in the forefront of state policy, then data systems are intensified to collect this type of information. Recently, state policy agendas have been shifting toward academic standards and teacher quality. Neither of these issues was prominent during the 1970s or early 1980s, and consequently, state data systems are underdeveloped in the very areas evaluators most need now.

Before demonstrating data needs and gaps in a specific state context, it is useful to outline what an ideal state system would encompass.[1]

1) The Environment of the Pupils: areas such as community and family characteristics, including parental education, occupation, income, and educational aspirations;

2) Pupil Capabilities and Difficulties: handicaps, language facility, motivation, etc.;

3) Resources that come to the School District: financial revenues and other resources such as volunteers;

4) Resources Allocated to Individual Schools: school site facilities, staff materials, and other purchased resources;

5) Education Goals: school district and school site goals, objectives, and curricular expectations;

6) Educational Opportunities: curriculum offerings, standards, teacher assignments and characteristics, and other school-related activities;

7) Pupil Participation: minutes spent learning subjects, courses taken, special programs pursued, extra-curricular participation;

8) Educational Outcomes: achievement, graduation, job placement, postsecondary enrollments, political participation.

No state approaches this comprehensive level of data, and rarely are the partial pieces of this schema linked for evaluative purposes. For example, school aid to local districts cannot be linked with the courses taught to see how many teachers are teaching in subject areas where they have inadequate academic preparation. Some states have detailed test data, but they cannot be linked to school site program offerings or to the specific courses that pupils have completed. The linkages between the data elements are as important as the existence of a comprehensive data base. Other states have good data on categorical programs, which are increasingly useful for monitoring local compliance with state regulations. Yet they lack new types of data essential to evaluate the more urgent comprehensive policy to comprehensive data base.

A good way to grasp these issues is to use a case study of one state. Therefore, the experience of California is analyzed below to demonstrate the type of revamping needed in state data systems in order to meet the new policy agenda being pushed by numerous state legislatures.

## CALIFORNIA'S EDUCATION DATA SYSTEM: A CASE STUDY

Like many states, California is badly lacking in the collection of systematic information regarding areas such as curricular trends, instructional

practices, quality and specialization of educational personnel, school climate, and expectations for student attainment. This chapter provides a start on such tasks by suggesting specific policy recommendations for closing data gaps, enhancing state and local decisions, coordinating "data streams," matching policy data to the reform agenda, and providing top-level data leadership.

The problems, however, are not merely data gaps, but the lack of centralization and coordination of the various data sources. Various independent data streams collected in a largely uncoordinated fashion are fed back to districts on a fragmented, piecemeal basis. Categorical programs, teacher preparation, university performance and pupil assessment all have independent data collection and analytical efforts that tend to divide education policy by artificial "program barriers" and levels of education. Although California has devoted considerable effort to cutting out redundant data requests, it has not developed a single coordinating point that matches policy issues with the various data streams. The California data system raises the same issues that are prevelant in all states.

California has several major data streams that follow traditional boundaries. For example, California Assessment Program (CAP) data are distributed to one group of local users and bilingual categorical compliance information to yet another group. The entire system has a top-down orientation with insufficient attention given to bottom-up user demands or access. The state testing program (CAP) is a notable exception. Otherwise, local users find it difficult (assuming that they want to use state data) to combine state data streams for a comprehensive review of school district performance. Similarly, Commission on Teacher Credentialing (CTC) data are managed separately from the State Department of Education data, preventing easy access or coordinated surveys for policymakers. Teacher certification was split off from the State Education Department in 1970. Since a surplus of teachers existed, little attention was give to data about quality or supply. State Department of Education financial apportionment data are useful for local financial modeling and projections, but not for policy analyses. The fiscal breakdowns around aggregate input categories (instruction, fixed charges, etc.) are not aligned

with the current reform issues stressing improvements in curricula and teacher quality.

The new education agenda, featuring higher "academic standards," is for the most part simply grafted on to state data priorities of the 1970s -- equalization and categorical program compliance. This approach presents a clear danger of using state testing programs such as CAP to collect more data than local school districts can respond to or will agree to provide. For example, CAP added 10 items in 1983 to probe "school climate," including discipline and homework. This increased state demand for data is resented by local district officials, largely because they do not find much of the state data easy to use or relevant to their particular problems. Moreover, local districts do not want to hear "bad news" from the state.

A share of the difficulty, however, resides with the local school districts who often do not use their own data (like needs assessments) for policymaking, either. A forthcoming Stanford study by Robert Calfee finds that much local information is collected with no connection to local policy or administrative decision-making processes and structures. The lack of local district use of state-collected data is a mutual problem of low motivation and high barriers at each policymaking level. Data are used for state compliance or audit purposes but less often for program evaluation or policy making.

Lack of local use of state data poses several problems. For example, classroom teachers must complete California Basic Education Data Systems (CBEDS), but they have little motivation to prepare accurate reports. Principals and teachers lack understanding of and commitment to CBEDS, and the state has difficulty coercing reluctant local educators to provide accurate numbers.[2] In most states, the State Department of Education merely notes that failure to submit complete CBEDS data will result in "incomplete data in federal and state reports." This may not be a great concern to local teachers and principals. State administrators reported in interviews that they could find DBEDS users no closer to the classroom than at the district level. Although the state collects school site performance data, accreditation is handled privately

and independently by the Western Association of Schools, and Colleges (WASC), largely in isolation of potentially useful state data. In addition, the Commission on Teacher Credentialing data are not even on computers, nor are they merged with either the Department of Education or WASC data. It is not known, for instance, how many are teaching in subject fields for which they do not have proper academic preparation. University of California (UC) and California State University (CSU) student performance data on placement tests and freshmen college grades are yet another separate, but relevant information stream. Feedback to local schools from UC and CSU is haphazard, piecemeal, and substantially less than maximally useful. Local educators seldom use university performance data as a basis for rethinking college preparation curriculum.

The State Education Department has made considerable progress in merging categorical compliance and CAP data -- the Consolidated Programs Description Data Base is a good example. California Assessment Program researchers have been able to construct profiles that enable one to identify particularly outstanding secondary schools on several dimensions. This promising start on combination an increased polciy relevance of data streams need to be intensified. Performance indicator systems can merge a number of these disparate data streams into one format for reporting to the public. Performance Indicators are discussed at length in Chapter 6.

The California performance indicator system (Appendix I) relies primarily on student questionnaires attached to the statewide testing program in order to present before-and-after measures of some parts of the new reform law. The statewide indicators presented are supplemented by a report on the same indicators for each high school. Data at the high school level also provide performance breakdowns by ethnic group. The data in this report happended to be collect prior ot the 1983 state reform and were deemed appropriate for use in monitoring reform impact after 1984.

Several states use the following criteria for deciding appropriate quality indicators. Indicators should:

- o Stem from overall goals for students;
- o Lead the instructional program in the right direction;
- o Be equitable -- that is, measure the progress of the below-average, and above-average students, and compare like student bodies;
- o Use data that are available and common across the state.

Here are several recommendations for state policymakers who want to improve their knowledge of the impact of their educational reforms.

RECOMMENDATION ONE

State education policy data should be supervised and coordinated by a single statewide office.

Perhaps the label "Information Czar" is not appropriate, but this concept does express the administrative essence. Some single office should be responsible for coordinating the combining data streams across state agencies (State Department of Education, teachers licensing agency, State Employment Department, university system, and community college system.). The Information Czar might have a board of directors composed of the state superintendent, university administrator, community college administrator, teacher credentialing agency, state finance director, legislative analyst, legislative audit committee, and local data users. This board should play a lead role in filling data gaps (see detailed recommendations below), eliminating low-use data sources in order to provide a holistic view of schools rather than individual categorical program reports. The cross-agency focus is especially appropriate given the reform focus on college preparation and measuring success at college for each high school and for various minority groups.

This Information Czar should have some specific, widely publicized criteria for approving collection of school performance measures. Some desirable characteristics of proposed measures are that:

1) They must reflect desired outcomes that are important.

2) The attribute to be measured must be susceptible to accurate measurement on some quantifiable basis.

3) People must be convinced of the validity and reliability of the measure as an indicator of the desired outcome.

4) Increases in the measure must be theoretically equally possible for every school, regardless of how well it is currently doing in the educational process, or of the kind of student body it serves.

RECOMMENDATION TWO

Major gaps in state education policy data should be filled as soon as possible.

1) In almost all states little data exists on middle schools. Little is known, for example, about how middle school tracks and course choices determine academic course work taken in the senior high school.

2) Typically, there is no integration between community colleges and elementary/secondary schools. Most states do not know how students from specific high schools perform in community colleges or what their grade point averages are in the first year. More states, however, have such data for the university system performance of high school students. Rarely is there any analysis of how graduates of specific high schools score on college academic placement tests. Since many community colleges are designing new data systems, integration with secondary school needs is particularly appropriate. This system may require a substantial amount of money for the community colleges.

3) Most of the state high school performance data focuses on the four-year-college-bound student. Many states are gathering more specific data on academic course-taking patterns but not on "life in the general or middle tracks" where fewer academic courses are taken. While categorical program data provide insight on the lowest achievers, these data are oriented to program

compliance rather than to state academic policy. Most states have achievement data on compensatory education pupils, but not on other student groups.

4) While most state reform laws require some new types of data, this collective effort should not be undertaken in isolation from existing data streams. Moreover, new data demands on localities to evaluate state reform can be combined with reductions in other data requests. States should look carefully to see if:

(a) some financial data are not used much, and

(b) whether some categorical-oriented data is no longer needed for monitoring because of less federal regulations under Chapter 1 of EICA.

5) The state assessment programs that test reading and math cannot be a freight train; that is, they cannot be used to carry vitally needed data to assess the impact of new state reforms. In California, the state test program (CAP) is already loaded with many new items such as school climate and course-taking patterns. The biggest reason for such current overuse is that CAP is the only data vehicle that reaches students. The state needs to consider a data student information sheet that is part os its basic data system along with finance and teacher characteristics. Any appreciable increase in data requirements attached to state testing programs probably will lead to increased local education agency resistance and lower data reliability.

6) Local school districts that report a large number of Limited English Proficiency (LEP) students must offer costly programs and seek waivers from the state for teachers who are not proficient in two languages. Consequently, while the Hispanic pupil population is growing dramatically, problems continue with the reliability and validity of local data regarding pupils with limited English-speaking ability. There are incentives for local school districts to underestimate LEP students and thereby also understate the need for teachers with dual-language capability and avoid state-mandated special programs.

7) There are serious shortcomings in existing data on the new policy dimensions regarding teacher quality. States need annual surveys of teachers working in subjects for which they have no credentials, supply-and-demand projections by subfield, and the probability that teachers in the reserve pool will return to the schools. Most states did not gather this data because of the teacher surplus from 1970 to 1982.

8) All states need a major effort to improve data on dropouts. Currently, states calculate attrition data, but not data on dropouts. As already mentioned, a pupil who dies is counted as a dropout in many states.

RECOMMENDATION THREE

State education data systems should emphasize and stimulate local user needs and adaptations.

States have been developing local users guides for their state testing programs. More needs to be done, however, to assist local school districts in assessing their academic standards and local curriculum policies. Selected test and pupil data at the local level should be on interlinked state and local microcomputers, and its accessibility by the state should be through telecommunication networking. Such a network would permit the state to extract its fair share of data and allow local districts to access state-developed software to analyze their own data locally. States could assist local performance reports by making it easier for local officials, through computer networks, to tap into state data bases (e.g., a distributed data base).

One promising Florida technique for enhancing local use is an annual report of school site performance that incorporates standard statewide data with locally devised categories (SEE APPENDIX II). The Florida approach was conceived in 1973 before the current stress on curriculum, academic content, and college performance. Consequently, it needs to be supplemented with numerous educational content and process variables reflective of the more recent reform efforts. The general concept has merit, however, and might provide a bottom-up performance data constituency. Florida reports are disseminated widely to

parents, newspapers, and citizens. The Florida system of financial rewards for the most informative school site reports has stimulated interest among local school districts. States could assist local performance reports by making it easier for local officials, through computer networks, to access state data bases directly.

A bottom-up performance data system will not be easy both to implement or to stimulate local enthusiasm for policymaking use. Much of the current locally generated needs assessment and test data are not used by local districts for their own policy decisions. Locally devised tests are used for individual pupil assessment, but much less often for local policymaking. Local school districts seldom collect systematic information on dropouts or postsecondary student outcomes. The state reform agenda, however, moves beyond a categorical compliance orientation to policies that local districts must be committed to in terms of academic time, content, processes, and standards. This new focus requires local commitment to data use, and will require an aggressive State Department of Educational technical assistance and field training effort. State data systems will never be "complete" or finished because as new policy areas arise a transition will take place. For instance, Mississippi is using output mesures based on the school effectiveness literature for its school accreditation.

Data can help make education systems undertake self-correcting changes and prevent overreaction to fads and media trends. A recent study by Denis Doyle of the American Enterprise Institute revealed that the federal government spends $9 million on education statistics, but $36 million on statistics on health, $137 million on labor, and $54 million on agriculture. Given these financial differences, it is not surprising that the nation knows more about its livestock than its students! Obviously, state governments cannot rely exclusively on an underfunded federal effort for financial support or for supplying comparative state performance data.

In conclusion, the redesign of state education data systems outlined above will cost much more money. Public education is a $140 billion a year

undertaking and therefore, warrants more state funds for information. In the future, local schools can computerize more performance data, but the use of data for education policymaking at the local level remains uncertain. State leadership should stress local use as well as generation of improved statewide education data. But a sole state focus on top-down data use by legislation never will generate the necessary enthusiasm at the local level to fill out the statwide forms accurately. Local school districts must have some stake in devising and using statewide data, which will be much harder to implement than new technical formats for statewide indicators. Recently, the U.S. Center for Statistics (formerly the National Center for Education Statistics) conducted a study of federal statistical needs in light of the changed state and local policy agenda. It appears that federal support will be growing for rethinking state bases. The more difficult intergovernmental relationship probably will be with local school districts who are reluctant to provide more data to states without special state funds for this purpose and without a clear connection of state data to local decisions.

## Chapter 5

## NOTES

*The writer would like to acknowledge the critical comments of Bill Burson and Vin Madden of the California State Department of Education. Comments also were made by James W. Guthrie and David Stern of the University of California, Berkeley.

1 This outline is derived from George Hall et al., "Alternatives for a National Data System," a report prepared for the U. S. Department of Education, December 1985.

2 Data in this paper on CBEDS is derived from Barbara Jean Sims' paper, "Who Makes Up CBEDS," Ms. Sims conducted 28 state and local interviews of educators. This paper was published in Education and Urban Society, vol 18, no.3, May 1986, pp. 359-368.

| Quality Indicator | STATEWIDE AVERAGE | STATEWIDE TARGETS | | |
|---|---|---|---|---|
| | 1983-84 | 1985-86 | 1987-88 | 1989-90 |
| **D. Performance of College-Bound Students** | | | | |
| **1. Scholastic Aptitude Test (SAT)** | | | | |
| Percent seniors taking | 39.0 % | * | * | * |
| Verbal | 421 | | | |
| Average score | | 428 | 436 | 444 |
| Percent of seniors scoring at least 450 | 15.5 % | 17.5 % | 21.0 % | 25.0 % |
| Mathematics | 476 | | | |
| Average score | | 481 | 489 | 496 |
| Percent of seniors scoring at least 500 | 16.5 % | 17.8 % | 19.1 % | 20.5 % |
| **2. Achievement Tests** | | | | |
| Test of Standard Written English | 42.6 | 43.0 | 43.5 | 44.0 |
| English Composition | 492 | 503 | 514 | 525 |
| Mathematics Level 1 | 517 | 528 | 539 | 550 |
| Mathematics Level 2 | 647 | 656 | 665 | 675 |
| American History | 501 | 509 | 517 | 525 |
| **3. Advanced Placement (AP)** | | | | |
| Number of scores 3 or better (per 100 seniors in school) | 9.5 | 12.0 | 15.0 | 18.0 |

NOTE: All data on this page are for 1983-84 school year; 1984-85 results are not yet available.

SOURCE: Califaornia State Department of Education, Performance Report for California Schools: 1985, Sacramento: California State Department of Education, 1985.

* Percent of takers for information only, no statewide targets have been established

| Quality Indicator | STATEWIDE AVERAGE | | STATEWIDE TARGETS | | |
|---|---|---|---|---|---|
| | 1983-84 | 1984-85 | 1985-86 | 1987-88 | 1989-90 |
| **A. Course Enrollments** | | | | | |
| **1. Mathematics** | | | | | |
| 3 or more years | 67 % | 74 % | 70 % | 73 % | 75 % |
| Advanced Mathematics (per 100 juniors and seniors enrolled) | 28 | 32 | 32 | 36 | 40 |
| **2. English** | | | | | |
| 4 or more years | 73 % | 86 % | 75 % | 78 % | 80 % |
| **3. Science** | | | | | |
| 3 or more years | 33 % | 36 % | 38 % | 42 % | 50 % |
| Chemistry (per 100 seniors in school) | 25 | 31 | 27 | 32 | 40 |
| Physics (per 100 seniors in school) | 10 | 12 | 16 | 21 | 25 |
| **4. History/Social Science** | | | | | |
| 4 or more years | 33 % | 37 % | 38 % | 42 % | 50 % |
| **5. Foreign Language** | | | | | |
| 3 or more years | 22 % | 22 % | 25 % | 29 % | 32 % |
| **6. Fine Arts** | | | | | |
| 1 or more years | 65 % | 67 % | 66 % | 68 % | 70 % |
| **7. Enrollment in a-f courses** | | | | | |
| Percentage of school enrollments in a-f courses | * | 38 % | 40 % | 45 % | 50 % |
| Percentage of graduates completing a-f requirements | ← | TO BE ANNOUNCED (Spring 1986) | | | → |
| **8. Units required for graduation** | ← | TO BE ANNOUNCED (Spring 1986) | | | → |
| **9. State Board Model Graduation Standards** | ← | TO BE ANNOUNCED (Spring 1986) | | | → |
| Percentage of graduates meeting standards | | | | | |

* Data first collected : 1984 85

SOURCE: California State Department of Education

| Quality Indicator | STATEWIDE AVERAGE | | STATEWIDE TARGETS* | | |
|---|---|---|---|---|---|
| | 1983-84 | 1984-85 | 1985-86 | 1987-88 | 1989-90 |
| **B. California Assessment Program (CAP) Scores** | | | | | |
| 1. Reading | | | | | |
| Percent correct score | 62.2 | 62.9 | 62.7 | 63.7 | 64.7 |
| Percent scoring | | | | | |
| above $Q_3$ | 25 % | 27 % | 26 % | 27 % | 28 % |
| above $Q_2$ | 50 % | 52 % | 52 % | 56 % | 60 % |
| above $Q_1$ | 75 % | 77 % | 76 % | 78 % | 80 % |
| 2. Mathematics | | | | | |
| Percent correct score | 67.4 | 68.3 | 67.9 | 68.9 | 69.8 |
| Percent scoring | | | | | |
| above $Q_3$ | 25 % | 29 % | 26 % | 27 % | 28 % |
| above $Q_2$ | 50 % | 48 % | 52 % | 56 % | 60 % |
| above $Q_1$ | 75 % | 77 % | 76 % | 78 % | 80 % |

* Targets are compared to 1983-84 standards.

| Quality Indicator | STATEWIDE AVERAGE | | STATEWIDE TARGETS** | | |
|---|---|---|---|---|---|
| | 1983-84 | 1984-85 | 1985-86 | 1987-88 | 1989-90 |
| **C. Dropout/Attendance** | | | | | |
| 1. Dropout rate (statewide averages and targets are attrition rates, grade 9 to graduation) | 29.3 % | 29.0 % | 28.3 % | 26.0 % | 23.5 % |
| 2. Attendance rate – percent of students in attendance on a specified day | ** | 93.7 % | 94.0 % | 94.5 % | 95.0 % |

** Data first collected in 1984-85

## APPENDIX II

### AN OVERVIEW OF A SCHOOL SITE ANNUAL PERFORMANCE REPORT FLORIDA

Whereas the statewide testing program provides the state with an early-warning system regarding minimum levels of student achievement, the Annual Performance Report is primarily for local client interests. This report would appear once each year, probably in early spring. The principal would be ultimately responsible for overseeing its production, but it should have sections reserved for exclusive use of the Parent Advisory Council, students (above the ninth grade), and staff. The report would be published in local newspapers, posted prominently in the school, and, most importantly, sent home to the parents or guardian of each student. The report would be the primary printed instrument by which clients could assess the effectiveness of their local school.

Proliferation of reporting forms and data collection efforts has long been a frustrating fact of life in both the private and public sectors. If Annual Performance Reports are well designed, they should help to reduce some of the burden. For the state, federal government and local school district as well as for the individual school site, the Annual Performance Report should be the primary data compilation instrument. The school district could aggregate the information it needed from school reports, and then pass them forward to the state. Rather than imposing an additional informational burden on local school personnel, the Annual Performance Report might well consolidate all other such efforts.

The contents of an Annual Performance Report should include topical categories and items such as those illustrated below:

**School Information**

Name, location, enrollment, age of building, number of classrooms, number of specialized rooms, school site size, state of repair, amount spent on maintenance in the last year and last decade, library volumes, etc.

**Staff Information**

Number of staff by category, proportion of various license classifications, age, sex, ethnic background, experience, degree levels, etc.

**Student Performance Information**

Intellectual performance: all results on student performance in standardized tests should be reported in terms of state-established minimum standards. Relative performance of different schools in the district should also be provided. Other performance information might also be included: student turnover rate; absenteeism; library circulation; performance of past students at next level of schooling (junior high, high school, college); etc.

**Areas of Strength**

Here the school can describe what it considers its unique or noteworthy characteristics. The purpose is to encourage every school to have one or more areas of particular specialization and competence, or to espouse a particular educational philosophy, or to employ a distinct methodology or approach. This section would inform parents about the tone or style of the school.

**Areas for Improvement**

This section would identify five areas in which a school needed improvement and would outline its plans regarding them. These problem areas might in some schools change over the years, but in others remain the same as the schools mounted a long-term improvement project. This section should encourage schools

Chapter 6

# STATE AND LOCAL PERFORMANCE

# INDICATORS AND COST/EFFECTIVENESS STUDIES

Michael W. Kirst
and
John Rogers

Evaluation of state education reforms should be designed with several different approaches, some of which include statewide indicators. These indicators can best be used in conjuction with other evaluation techniques that can probe deeper inside schools and classes. The dangers of performance indicators are highlighted by the controversy surrounding the "wall chart" first released by former U.S. Secretary of Eduction T. Bell in January 1984, containing comparisons of test scores and high school graduation rates of the 50 states. The statistical comparisons for Alabama, Alaska, Arizona, and Arkansas are presented below. Note the limited amount of data on state social and economic context that might account for much of the interestate difference. All of these states use ACT, but most national media only report SAT results.

Table 1.

State Education Statistics,

State Education Performance Outcomes, Resource Inputs and Population

Characteristics, 1972-1982

to be self-critical, to establish specific goals and to report on subsequent progress.

Parent, Teacher and Student Assessment of School Performance

Responsible parents, teachers and students should be permitted an uncensored opportunity to assess school performance. This section would permit various school constituencies to express their opinions of school success or failure with respect to such matters as instruction, curriculum development, racial relations, student participation in decision making, drug abuse, etc.

---

Source: James Guthrie, University of California Berkely, as cited in Larry Pierce et. al., State School Journal Alternatives (Eugene: University of Oregon, 1975.)

## Performance Outcomes

| | | American College Testing Program Scores for 28 States | | Scholastic Aptitude Test (SAT) Scores for 22 States | | Graduation Rate | |
|---|---|---|---|---|---|---|---|
| | | Act Score | Rank Out of 28 | SAT Score | Rank Out of 22 | Percent | Rank Out of 51 |
| Alabama | 1982 | 17.2 | 26.0 | — | — | 67.1% | 43.0 |
| | 1972 | 18.3 | 25.5 | — | — | 65.4% | 48.0 |
| Alaska | 1982 | 18.7 | 13.5 | — | — | 71.0% | 38.0 |
| | 1972 | 19.6 | 15.5 | — | — | 79.3% | 25.0 |
| Arizona | 1982 | 18.7 | 13.5 | — | — | 72.4% | 33.0 |
| | 1972 | 19.3 | 17.5 | — | — | 71.8% | 37.0 |
| Arkansas | 1982 | 17.7 | 20.0 | — | — | 74.7% | 30.0 |
| | 1972 | 18.6 | 22.0 | — | — | 68.9% | 45.0 |

| | | Resource Inputs | | | | Population Characteristics | | |
|---|---|---|---|---|---|---|---|---|
| | | Current Expenditures Per pupil | | Expenditures as Percent of Income Per Capita | | | | |
| | | Expenditure | Rank Out of 51 | Percent | Rank Out of 51 | Percent Poverty Ages 5-17 | Percent Minority Enrollment | Percent Handicapped Enrollment |
| Alabama | 1982 | $1,835 | 46.5 | 21.2% | 32.0 | 23.1 | 37.0 | 9.5 |
| | 1972 | $ 843 | 50.0 | 15.9% | 50.0 | 29.5 | 34.5 | — |
| Alaska | 1982 | $5,369 | 1.0 | 33.0% | 1.0 | 11.4 | 22.5 | 8.2 |
| | 1972 | $1,441 | 2.0 | 27.3% | 3.0 | 14.6 | 16.3 | — |
| Arizona | 1982 | $2,305 | 26.0 | 22.7% | 21.0 | 15.8 | 29.8 | 9.3 |
| | 1972 | $ 911 | 22.0 | 21.3% | 17.0 | 17.5 | 27.2 | — |
| Arkansas | 1982 | $1,713 | 50.0 | 20.2% | 40.0 | 22.7 | 30.2 | 9.7 |
| | 1972 | $ 801 | 49.0 | 18.4% | 42.0 | 31.6 | 29.4 | — |



1-10

Source: U.S. Department of Education, National Center for Educational Statistics, 1984.

These comparisons led to a storm of criticism that there was a lack of context and controls in interpreting these "statewide" outcomes. SAT scores, for example, were not adjusted for the percent of high school students in each state that took the test. In states with very few SAT test takers, results are usually higher because primarily the better students take SAT. States with a very small number of SAT test takers were compared with others that had a majority of their students taking the test. State officials also assailed the wall chart for its superficial nature and the lack of valid data in areas such as dropout rates.

Ironically, the misleading nature of the wall chart stimulated the Council of Chief State School Officers (CCSSO) to create its own sophisticated system of state-by-state comparisons. In 1985 CCSSO decided to take an active role both in determining what indicators should be used to compare states and in establishing a center to coordinate nationwide assessment activities. CCSSO may work on a standardized national achievement test, but in the short run it will implement other interstate indicators. In a growing number of states, including Florida and Virginia, officials are using NAEP to compare their own pupil achievement scores and similarly, hope to persuade other states to use this national test. NAEP has never provided state comparisons, and the numerous tests that states use now do not permit such rankings.

This movement for interstate indicators has reinforced initiatives in several states to construct comparisons of school system performance before and after reform. State performance indicator systems are typically expensive to develop and take a considerable time to mount. One of these efforts is in South Carolina, where state officials have conducted public opinion polls of citizens and teachers to gauge their perceptions of reform. Another one in California provides not only statewide indicators of course-taking patterns, test, attendance, college performance, and so on, but also a performance report for each school (see Chapter 5.) California divides its schools into five comparison groups based on the socioeconomic status of the pupils. The "percentile in group" reported in Table 2 is the ranking within the appropriate

comparison band. Beverly Hills and Los Angeles are in quite different comparisons groups.

The California assessment system is noteworthy because it splits test scores by quartiles in order to see how the various tracks (academic, general, vocational, and remedial) are performing. A number of high or low scorers can affect the average score in a manner that will exaggerate performance for the middle tracks. Some schools do very well with the college bound group but quite poorly with other students. The use of an average test core for a school masks this differential performance among different types of students. Much of the recent state reform is focused on the general and vocational tracks, and these breakdowns of attainment by quartiles can provide some insight on the progress of these two tracks (see chart in Table 2).

## INDICATORS OF SUBJECT MATTER ATTAINMENT

A recent study by the National Research Council (NRC), Indicators of Precollege Education in Science and Math (1985), provides a conceptual diagram for analyzing the input/output relationships in a specific subject area.[1]

Education System

| Inputs | Process | Outcomes |
|---|---|---|
| Teachers ----------> | Instructional ----> | Student achievement |
| Quality | Time | |
| Quantity (supply & demand) | Course enrollments | |
| | Homework | |
| Curriculum Content | Allocated vs. actual | |
| Opportunity to learn | instruction time | |
| (exposure to specific content) | | |
| Textbooks and Courses | | |

1

C
O
S
CO

2

1984-85 GRADE 12 ENROLLMENT: 310
1984-85 GRADE 12 TESTED: 299
1984-85 PERFORMANCE GROUP: D
1983-84 PERFORMANCE GROUP: D

9

ETHNIC DISTRIBUTION BY GRADE LEVEL, 1984-85*

| ETHNIC GROUP | GRADE 12 NUMBER | GRADE 12 PERCENT | GRADE 11 NUMBER | GRADE 11 PERCENT | GRADE 10 NUMBER | GRADE 10 PERCENT | GRADE 9 NUMBER | GRADE 9 PERCENT | SCHOOL NUMBER | SCHOOL PERCENT |
|---|---|---|---|---|---|---|---|---|---|---|
| WHITE | 209 | 67.4 | 229 | 71.8 | 283 | 73.9 | 322 | 71.2 | 1,043 | 71.3 |
| BLACK | 11 | 3.5 | 8 | 2.5 | 6 | 1.5 | 11 | 2.4 | 36 | 2.4 |
| HISPANIC | 22 | 7.1 | 21 | 6.6 | 24 | 6.1 | 41 | 9.1 | 108 | 7.3 |
| ASIAN/FILIPINO | 68 | 21.9 | 59 | 18.4 | 69 | 17.7 | 77 | 17.0 | 273 | 18.5 |
| AMERICAN INDIAN | 0 | 0.0 | 2 | 0.6 | 3 | 0.8 | 1 | 0.2 | 6 | 0.4 |
| ALL STUDENTS | 310 | 100.0 | 319 | 100.0 | 391 | 100.0 | 452 | 100.0 | 1,472 | 100.0 |

10

ENROLLMENT RATES OF ETHNIC GROUPS IN SELECTED COURSES FOR GRADE 12, 1984-85*

| ETHNIC GROUP | PHYSICS NUMBER | PHYSICS RATE** | CHEMISTRY NUMBER | CHEMISTRY RATE** | ADVANCED MATHEMATICS NUMBER | ADVANCED MATHEMATICS RATE*** |
|---|---|---|---|---|---|---|
| WHITE | 23 | 11.0 | 106 | 50.7 | 70 | 16.0 |
| BLACK | 2 | 18.2 | 1 | 9.1 | 3 | 15.8 |
| HISPANIC | 0 | 0.0 | 5 | 22.7 | 7 | 16.3 |
| ASIAN/FILIPINO | 17 | 25.0 | 36 | 52.9 | 43 | 33.9 |
| AMERICAN INDIAN | 0 | . | 0 | . | 0 | 0.0 |
| ALL STUDENTS | 42 | 13.5 | 148 | 47.7 | 123 | 19.6 |

* ITEM IS FOR THE PURPOSE OF PROVIDING ADDITIONAL INFORMATION FOR THE SCHOOL AND IS NOT USED AS A QUALITY INDICATOR.

** PHYSICS AND CHEMISTRY RATES ARE BASED ON PER 100 SENIORS ENROLLED.

*** ADVANCED MATHEMATICS RATE IS BASED ON PER 100 JUNIORS AND SENIORS ENROLLED.

SOURCE: California State Department of Education, Office of Special Studies and Evaluation Reports, 1985.

# TABLE 2

INFORMATION SERVICES — HIGH SCHOOL PERFORMANCE REPORT SUMMARY, 1984-85 — DATE: 09/09/85

**1**

C
O
S
CO

**2**

1984-85 GRADE 12 ENROLLMENT: 310
1984-85 GRADE 12 TESTED: 299
1984-85 PERFORMANCE GROUP: 8
1983-84 PERFORMANCE GROUP: 8

**3** COURSE ENROLLMENT, 1984-85

| COURSE | PERCENT | CHG. FROM 1983-84 | PERCENTILE IN GROUP |
|---|---|---|---|
| 3+ YRS MATH | 75.7 | 7.8 | 75 |
| ADVANCED MATH | 33.7 | 6.1 | 66 |
| 4+ YRS ENGLISH | 90.0 | 9.7 | 78 |
| 3+ YRS SCIENCE | 38.7 | -1.5 | 76 |
| CHEMISTRY | 47.7 | 6.4 | 89 |
| PHYSICS | 13.5 | -2.2 | 69 |
| 4+ YRS SOCIAL SCIENCES | 22.9 | 8.2 | 37 |
| 3+ YRS FOREIGN LANGUAGES | 41.0 | 7.7 | 98 |
| 1+ YR FINE ARTS | 54.2 | -7.3 | 18 |
| PERCENT IN A-F COURSES | 35.9 | NOT APP. | 30 |

**4** CALIFORNIA ASSESSMENT PROGRAM, 1984-85

| SUBJECT | PERCENT | CHG. FROM 1983-84 | PERCENTILE IN GROUP |
|---|---|---|---|
| READING SKILLS | | | |
| PERCENT CORRECT | 63.3 | 0.6 | 35 |
| PERCENT SCORING ABOVE Q3 | 26% | -2 | 40 |
| PERCENT SCORING ABOVE Q2 | 52% | -1 | 35 |
| PERCENT SCORING ABOVE Q1 | 79% | 0 | 49 |
| MATHEMATICS SKILLS | | | |
| PERCENT CORRECT | 70.2 | 1.1 | 63 |
| PERCENT SCORING ABOVE Q3 | 30% | 2 | 73 |
| PERCENT SCORING ABOVE Q2 | 52% | 0 | 69 |
| PERCENT SCORING ABOVE Q1 | 81% | 2 | 61 |

**5** PERFORMANCE OF COLLEGE BOUND STUDENTS

| SCHOLASTIC APTITUDE TEST | NO. OF TESTS | 1983-84 RESULTS | PERCENTILE IN GROUP |
|---|---|---|---|
| SENIORS TESTED | 151 | 49.5% | NOT APP. |
| VERBAL SKILLS | | | |
| AVERAGE TEST SCORE | 151 | 418 | 33 |
| TESTS WITH SCORE>=450 | 56 | 18.4% | 88 |
| TESTS WITH SCORE>=500* | | 11.8% | NOT AVAIL. |
| TESTS WITH SCORE>=600* | 6 | 2.0% | NOT AVAIL. |
| MATHEMATICS SKILLS | | | |
| AVERAGE TEST SCORE | 151 | 494 | 50 |
| TESTS WITH SCORE>=500 | 68 | 22.3% | 95 |
| TESTS WITH SCORE>=550* | 39 | 12.8% | NOT AVAIL. |
| TESTS WITH SCORE>=600* | 24 | 7.9% | NOT AVAIL. |
| ACHIEVEMENT TESTS | | | |
| AVERAGE STD. WRITTEN ENGLISH | 151 | 41.9 | 29 |
| AVERAGE ENGLISH COMPOSITION | 50 | 482 | 34 |
| AVERAGE MATH LEVEL 1 | 49 | 541 | 69 |
| AVERAGE MATH LEVEL 2 | 6 | 660 | 68 |
| AVERAGE AMERICAN HISTORY | 20 | 459 | 18 |
| ADVANCED PLACEMENT TEST | | | |
| TESTS WITH SCORE OF 3/BETTER | 16 | 5.2% | 57 |

**6**

| | | | |
|---|---|---|---|
| ATTENDANCE RATE (PERCENT) | 100.0 | NOT APP. | 88 |

**7** LANGUAGE PROFICIENCY AND MOBILITY ON TEST SCORES, 1984-85*

| LANGUAGE PROFICIENCY AND MOBILITY | PERCENT OF STUDENTS | PERCENT CORRECT FOR: READING | MATHEMATICS |
|---|---|---|---|
| LANGUAGE, ALL | 100.0 | 63.3 | 70.2 |
| LIMITED ENGLISH | 5.0 | 38.1 | 61.4 |
| FLUENT ENGLISH & ENGLISH ONLY | 95.0 | 64.8 | 70.8 |
| MOBILITY (GRADES 10-12) | 17.0 | 63.4 | 73.7 |

**8** COLLEGE PERFORMANCE, 1982-83*

| FRESHMAN PERFORMANCE | NO. OF STUDENTS | AVG. GPA | DIFF. OF GPA |
|---|---|---|---|
| FRESHMAN GRADE POINT AVERAGE (GPA): | | | |
| AT UNIVERSITY OF CALIFORNIA (UC) | 26 | 2.92 | 0.03 |
| AT CALIFORNIA STATE UNIVERSITY (CSU) | 30 | 2.68 | 0.17 |

* ITEM IS FOR THE PURPOSE OF PROVIDING ADDITIONAL INFORMATION FOR THE SCHOOL AND IS NOT USED AS A QUALITY INDICATOR.

NRC examined the best available measures for each of these variables and summarized the available data. The council found little consensus and scant data about teacher quality and some other variables, including actual instructional time and exposure to specific content. NRC did not include important variables, such as teacher enthusiasm and instructional technique, that would help students to be interested, challenged, and to comprehend complex scientific material. But a number of these indicators were available and not too difficult to collect. Several relationships were affirmed, including the significant relationship between instructional time and course enrollments with student outcomes. A sample of their findings is presented below.

FINDINGS

Instructional Time and Student Learning:

o The amount of time given to the study of a subject is consistently correlated with student performance as measured by achievement tests, at the elementary school as well as the secondary school level. The key issue in achievement is whether a student has been able to study or has been exposed to a particular content area. For example, students rarely learn physics or French if they have not studied this subject in school. The interaction of time and exposure to content is the crucial relationship.

o Time spent on homework also increases with student achievement. A key issue is whether the student receives feedback from the teacher concerning his homework. If not, there is much less impact on achievement.

Measuring Instructional Time: Elementary School

o For elementary schools, not enough data are available to discern clear trends over the last 20 years with respect to amount of instructional time spent on mathematics and science. On average, about 45 minutes a day are spent on mathematics and 20 minutes on science. Existing information, however, points to great variability from class to class in the amount of time given to instruction in general and to each academic area specifically. At any rate, the scant

attention to science at the elementary level has become a policy issue in several states.

Measuring Instructional Time: High School

o The average high school senior graduating in the early 1980s has taken about 2-3/4 years of mathematics and 2-1/4 years of science during grades 9-12.

o Compared with 20 years ago, average enrollments of high school students in science have declined. While this trend now appears to be reversing, enrollments have not returned to the level of the early 1960s.

o High school enrollments in mathematics have increased over the last decade by about a semester.

o College-bound students are taking more mathematics and physical science courses in secondary school than they did 10 years ago, and the increases have been continuous throughout that period. The gap in enrollment between males and females in advanced mathematics courses is narrowing.

o A number of problems attend enrollment data currently available: uncertainties generated by using self-reports, differences in questions and method from survey to survey, and ambiguities created by similar course titles in mathematics that refer to different content or different levels of instruction.

Critics of statewide indicators contend that they confuse local token compliance with state policy and easily observable standards (e.g., every student takes four years of English) with basic changes in teacher and students classroom performance. Districts conform to the letter of top-down directives but seldom go beyond token compliance unless they are internally motivated to do so. Quantity becomes confused with quality as state policymakers seek simplistic bottom-line numbers for evaluative purposes.

A locally devised performance indicator system, however, can help overcome some of these problems. While school climate, homework, and writing cannot be

measured well at the state level, local student surveys, teacher reports, and interviews can provide insight on several of these areas. Students can be asked whether they feel safe at school, writing samples can be collected and analyzed, and teachers can be interviewed about their homework policies. It is difficult to derive summary numbers from such data, but trends, effective practices, and analysis can be enlightening. The California Commission on the Teaching Profession has recommended that each school prepare a biennial "Index of Conditions." The index would include:

(1) class size and teaching load;
(2) teacher assignment outside area of competence;
(3) time spent by teachers on nonteaching tasks;
(4) inefficiency and currency of texts and teaching materials;
(5) availability of substitutes;
(6) school order, climate for learning, and behavior problems;
(7) adequacy, safety and cleanliness of school facilities;
(8) adequacy of teacher evaluation and opportunities for professional improvement.

In sum, performance indicators can be an important part of an overall system to evaluate state reforms. They can provide statewide trend data on variables of high interest to reformers as well as implications for improved pupil performance. Many indicators are better than a few, and local supplements to statewide data can give a more complete picture. Even the best indicators, however, offer only a partial overview of reform and may confuse surface implementation with classroom reality. Several indicators such as dropouts, school climate, writing, and homework are difficult to devise with high standards of reliability and validity.

## EVALUATING COSTS AND ASSESSING COST EFFECTIVENESS

While several states have begun to develop performance indicators, very few have included cost studies as part of their effort. Yet, how can legislators judge performance and its relationship to school resource allocations without a

good grasp of cost? Performance outcomes can be generated at an extremely high state cost. On the other hand, there may be methods to obtain desired results at relatively low cost. Evaluators have developed an analytical scheme called cost/effectiveness to dissect these cost/quality relationships. This issue is especially important because most recent state reforms were not costed out in detail nor did they allocate a specific amount to accomplish the legislative objectives. Rather, states created a list of reforms and often appropriated a lump sum appropriation to cover most of them. State legislators did not know how much a career ladder program would cost or the impact of increased state graduation standards on local curriculum trade-offs between academic subjects or vocational subjects.

To understand what has happened to reform dollars, attention must be focused on the changes in the financing of education created by the multitude of reform bills enacted over the past three years.[2] These changes vary between states and between programs within states. Some states, such as Arkansas, Florida, South Carolina, Tennessee, and Texas, have raised states taxes specifically to finance large education reform programs. Other states, such as California, Georgia, and Kentucky, have passed broad-based reform programs without increasing state taxes. Furthermore, within most states' financing plans, some programs are funded entirely by the state, some receive partial state support, and still other must be funded primarily by local districts.

Given the diversity of funding patterns for reform programs, it is not surprising that within each state, a multiplicity of answers exist to the crucial questions of:

1) Which reform programs receive financial support?

2) Who provides the money?

3) How much fiscal support is given?

4) Who receives the funds?

5) What happens to these funds?

A first area of evaluation should focus on questions 1 through 3 above; namely, which programs, school services, employee group, or expenditures categories have received funding and how much fiscal support have they received? If some programs are underfunded, then the legislature will need to consider additional funds. In Texas, the legislature has not fully funded the teacher career ladder, causing significant local property tax increases. This property tax impact leads into a second area of potential inquiry that resemble studies in financing equity, a major topic of the 1970s. Legislators need to know whether the recent reform efforts are reinforcing or in conflict with the earlier programs to equalize funding among rich and poor districts.

While most reform bills identify specific programs to be implemented, few programs receive separte funding from the state. (Florida's writing program and the Texas early childhood effort are exceptions.) In general, states have expected districts and schools to use increases in general state aid and local property taxes to support state education reform objectives. Given that local districts have had a great deal of discretion over funding decisions, the question becomes: which reform objectives ultimately received fiscal support (through state and/or local funds), and what local funding strategies were developed to address the new objectives?

Did districts fund low-cost programs first or those with high political visibility (such as increasing teacher salaries)? Did districts try to implement state intent with only partial funding; if so, which parts did they choose to implement?

Related to these issues is the question of how funding decisions were affected by the structural design of the state aid. Did local districts pay more heed to reform objectives funded through state categorical programs? Were local decisions affected differently by state flat grants than by aid through state equalization formulas? In short, evaluation is needed to examine what factors influenced local spending decisions. With this knowledge states can

know better the local impact of using a matching grant, incentive grants, pilot projects, or fully funded formulas.[3] What type of state funding mechanism elicits what kind of local spending reaction is the essential issue.

A crucial task for evaluators should be to determine what local response or mix of responses is most common for a given state reform program. For example, there are three possible local responses to any state grant. First, a state grant can increase the amount of local educational services above what otherwise would exist. Second, a state grant can help reduce support for a local educational service, thereby allowing the district to reallocate funding to other services. For example, a district already committed to providing high salaries to first-year teachers simply might use state funds designated for this purpose for other fiscal needs. Finally, the state grant can be used to reduce local property taxes or other local revenue sources and not result in any net increase in local education spending.

## Issues of Equity

The complex funding plans employed in the 1970s in many states to insure finance equity in education have been clouded by the implentation of a variety of reform programs. One impact is caused by the practice of funding certain state reform programs on a voluntary rather than a mandated basis. The voluntary programs employ state incentives to attract the participation of districts, schools, or teachers. A crucial equity issue that needs further research is whether these voluntary programs have attracted a disproportionate number of high property-wealth districts or districts in particular geographical locations. Are the fiscally poor school districts unable to take advantage of the voluntary or incentive programs?[4]

## Cost-Effectiveness

A third type of evaluation would employ cost-effectiveness. By studying the costs as well as evaluating the effectiveness of a number of interventions aimed at a commom goal, the policymakers gains an understanding of the relative worth of each intervention. This coupling of standard evaluation procedures with cost

analysis enables a decision-maker to choose the program or set of programs that provides the best educational results for any given state appropriation.[5] For example, data in California suggest that the state has purchased on the average only an extra 4 minutes for the school day from the hundreds of millions of dollars it has spent on longer school day incentives. While state funds led some districts to restore the sixth period of high school, more often there was little impact. On the other hand, California mentor teaching and high school counseling programs led to substantial changes in school procedures at relatively low costs. Given these general figures of cost and program impact, this suggests that California legislators should divert funds from the longer school day incentives to the mentor teacher or other more cost/effective programs.

Although this illustration demonstrates the power of combining cost analysis and evaluation, such a cursory approach is not precise enough to enable a policymaker to choose among a number of similar cost-effectiveness alternatives. What is needed is a more systematic approach to combining the evaluation of the program's impact with an assessment of its costs. One such approach is cost-effectiveness analysis (CE). CE evaluates a series of alternative programs, all of which seek a common outcome by comparing the ratios between each program's costs to each program's effectiveness.

This approach can best be explained through an illustration. In a hypothetical study, three vocabulary enhancement programs are offered to separate classes of 11th graders with the intent of improving the students' SAT scores. Program A gives the students two vocabulary textbooks to study at a cost of $30 per student. Program B uses the same textbooks but also employs a teacher to review the words once a week for five weeks at a cost of $100 per student. Program C computers to teach the students vocabulary words. The cost of the computer use, the computer programs, and teacher supervision totals $200 per student.

Table 3 depicts the costs of each program along with the average gain

TABLE 3

Hypothetical cost-effectiveness results for vocabulary enhancement program:

| Program | Cost Per pupil | Improvement Points | CE Ratio |
|---|---|---|---|
| A | $ 30 | 5 | $6 per point |
| B | $100 | 33 | $3 per point |
| C | $200 | 40 | $5 per point |

on SAT scores of students enrolled in the program. The third column displays the cost effectiveness ratio or the cost per student for a one-point improvement. The hypothetical results show that Program A cost the most for every point of improvement, while Program B is the most cost-effectiveness. Significantly, the most effective program when cost per student are ignored (Program C) is ranked second when cost per point are taken into account. Hense, a greater state or local investment in Program B is the perferred policy.

One limitation of cost-effectiveness analysis is that it can be used only to compare programs with similar or identical goals. When a comparison of programs with disparate goals is needed, a second type of analysis, cost-benefit (CB) should be employed. CB analysis evaluates a series of alternative programs by comparing the ratios between the program's costs to the monetary value of the program's benefits. This approach presupposes that a program's results can be quantified in monetary terms. CB analysis, for example, could compare two vocational education programs by looking at the jobs (and salaries) of each program. A problem with this approach is the difficulty of determining monetary gains (such as how much money students will earn). Because the evaluation procedures discussed in the previous chapters do not assess programs in this way, this section does not look at CB analysis in depth.

Assessing Costs

Research on the cost-effectiveness of the various reform programs should begin by employing the evaluation procedures outlined in Chapter 2. Analysis of costs, however, will require additional assessments.

Any analysis of the program's costs begins with two assumptions: (1) the cost of any program is the value of all the resources that the program uses, and (2) costs refer to the least expensive set of alternatives that will satisfy a given need.

Given these two assumptions, the question arises: Why can't the costs of a program be determined simply by looking at the budget expenditures for each education program or state reform? While such a straight-forward approach would facilitate conducting the analysis, there are four reasons why the budget fails to accurately predict a program's true costs. First, budgets often do not include all resources used in a program. For example, a budget does not account for contributed resources or other unpaid inputs such as a free building or volunteer classroom aides. Second, budgets generally do not include costs of resources such as buildings or heavy equipment that are used in the state reform, but have been paid for in past budget years. Third, budgets may lump funding of a number of different stte interventions together, making it difficult to distill the separate costs of each program. Finally, budgets represent plans for how resources should be distributed and often do not reflect the actual allocation of funds once the program gets underway.

What is needed is an approach to studying costs that identifies the start-up, medium, and long-term cost, as well as the hidden costs of the programs. Unfortunately, little analytic work exists on the true costs of the different state reform programs.

Research into costs generally can be modeled after the "ingredients method." Levin outlines five categories of cost inputs in educational programs that need to be assessed:[6] (1) personnel, (2) facilities, (3) equipment, (4) client inputs, and (5) supplies and other inputs.

Following are guidelines for assessing the costs of each of these ingredients as an aid to the direction cost research needs to go.

## THE INGREDIENTS METHOD

### Personnel

The personnel costs of reform programs often are the major cost input. Salaries of teachers, administrators, and support staff constitute the chief facet of such expenditures. When nonsalary benefits are given to personnel as part of a reform program's incentives, those benefits must be counted as well. For example, fringe benefits such as changes in medical insurance, a pension plan, or providing transportation to teachers all must be included. In programs that assign additional tasks to existing personnel, one must figure both the cost of these employees' time and those of additional personnel who might be needed to fill vacated positions. Additionally, when programs use volunteer workers, one must estimate the market value of the donated services and add this estimate to the total personnel expenditures.

### Facilities

The cost of any program must include fees for using additional facilities or for using existing ones more. This assessment can be relatively straightforward: one determines the annual cost of leasing the additional space, for example. Often, however, this approach is unfeasible. It is difficult, for instance, to determine the market value for a school building that is idle during the summer. A second approach would be to calculate the annual depreciation on the additional space.

### Equipment

Once again, expenditures for any equipment needed for a project must be considered. The analysis must determine how much the equipment would be to rent for a year or the depreciation on it over a year. For example, planners of a

program that put a computer in every school in the district should figure what the computers would cost for that one year, not the total purchase price.

Client Inputs

Any financial burden that students or their families must bear as part of the program also must be assessed as part of its costs. For example, transportation charges incurred by families must be figured as part of the costs of an enrichment program that offers summer school but no transportation services.

Supplies

It is often difficult to accurately estimate supplies for a project because of the variety used, each having a different price. Because supplies represent such a small percentage of total expenditures, only rough estimates can be made of their costs without affecting the accuracy of the overall analysis.

## USING COST-EFFECTIVENESS ANALYSIS

The ingredients method can help assess the cost of the different reform programs in place. Once the expenditures are established, the cost-effectiveness can be compared of disparate reform strategies such as improving instruction, reducing class size, lengthening school days, providing computer-assisted instruction, or establishing peer tutoring. A recent study shows, for example, that both peer tutoring and computer-assisted instruction are more effective in improving basic skills and less costly than smaller classes and longer school days. This early study points to the potential power of cost-effectiveness analysis to provide the policymaker a better understanding of which direction future reform efforts should take.[7]

At the same time, it needs to be emphasized that there are limitations on cost-effectiveness analysis. Due to the margin of error that exists both in evaluation of programs and in assessment of costs, slight differences in cost-effectiveness ratios should not be considered statistically significant.

At best, such analysis estimates the relationship between costs and program achievement. The analysis should be used as one source of information in the decision-making process.

## Chapter 6

## NOTES

1. Senta A. Raisen and Lyle V. Jones, Indicators of Precollege Education in Science and Math (Washington: National Academy Press, 1985).

2. In outlining this first area of research, we draw heavily on Allan Oden, "A School Finance Agenda for an Era of Education Reform." Paper prepared for the Institute for Educational Finance and Governance, Stanford University, 1985.

3. Mun Tsang and Henry Levin, "The Impact of Intergovernmental Grants on Education Expenditures," Review of Educational Research, Fall 1983: 53, no. 3: 329-367.

4. See, generally, Jay Chambers and Thomas Parrish, "The Issue of Adequacy in Financing of Public Education: How Much Is Enough?" Stanford University: the Institute for Educational Finance and Governance, IFG 82-A19, 1982.

5. This cost analysis section is an adaptation of Henery Levin's Cost Effectiveness: A Primer. (Sage Publications: Beverly Hills, 1983).

6. Levin, Cost Effectiveness.

7. H. M. Levin, G. V. Glass, G. R. Meister, "Cost Effectiveness of Four Educational Interventions" (Stanford: IFG Project Report No. 84-A11, May 1984).

Chapter 7

# HOW WELL DO U.S. SCHOOLS PERFORM?*

Michael W. Kirst

## Author's Introduction

The recent state reform movement has been in part stimulated by comparisons with other nations. Part of the reform rationale was to enhance the skills and abilities of the U. S. Labor force in order to better compete economically in a world economy. Some of the concern about U.S. education has been stimulated by international comparisons on student achievement scores, course-taking patterns, and school attendance. Unfortunately, many of these comparisons are between apples and oranges. They result in misleading implications for major policy directions. The Japanese and German eduction systems, for example, contain fundamentally different traditions, structures, and policy dimensions from the U.S. One has to be exceedingly careful in evaluating and comparing these three systems and in reaching policy conclusions. Accordingly, this chapter is designed to examine the different policy objectives of various nations, and to provide some ways to evaluate pupil performance on a more standardized basis than is commonly done in the mass media.

The egalitarian goals of the American system distinguish it from the schools systems of the European democracies. About 75 percent of U.S. students graduate from high school, and some 44 percent go on to postsecondary education. In most other Western nations, students are diverted into vocational and technical programs at age 14 or 15, and only 15 to 30 percent are graduated from a secondary school. Considering the large percentage of teenagers the U.S. schools enroll, the level of attainment of these students is surprisingly high.

---

* From Who Controls Our School? by Michael W. Kirst, W. H. Freeman and Company.
Copyright (c) 1985. Reproduced by permission of the publisher.

Our school system has grown steadily more egalitarian. As recently as 1940, fewer than 50 percent of the pupils in this country completed high school. by 1984, 75 percent did. Conscious and determined effort over the last 20 years has increased access to higher education among those least represented in the past--the minority groups. While college enrollment of white students grew slightly, the proprtion of Blacks in college more than doubled in the same period, growing between 1965 and 1976 from 5 to 11 percent. In 1981, the percentage of black high school graduates who went on to college exceeded that of whites for the first time. That many of these students come from lower socioeconomic groups makes this achievement all the more remarkable. Recent sociological studies from the Soviet Union, by contrast, indicate that an unexpectedly high share of the places in most of its elite postsecondary institutions go to children of white collar workers.

Lately, however, the headlines have warned us that our schools are not competitive with those of our economic rivals West Germany and Japan, and that current graduates of our secondary schools canot match the records set by their predecessors as recently as 1970. Now, when the performance of our schools is a source of widespread dissatisfaction, we need to consider carefully how we measure that performance.

## Comparing Apples and Oranges

One was to measure performance is to test "outcomes" -- the skills or knowledge gained by students -- and to compare what students learn in the United States with what students at the same level learn in West Germany or Japan. The openness of the American system, however--its power to retain students up to the age of 17 or 18-- makes it difficult to draw valid comparisons of outcomes between this and other systems. As Table 1 shows, the U.S. ranks first in percentage of 15- to 24-year-olds enrolled in school.

TABLE 1

**Enrollment Rates for Students Aged 15–18 and 20–24**

| | Age 15–18 | | | Age 20–24 | | |
|---|---|---|---|---|---|---|
| Country | Year | Percent Enrolled | Ranking | Year | Percent Enrolled | Ranking |
| Austria | 1969 | 31.9% | 14 | 1972 | 14.8% | 12 |
| Belgium | 1966 | 54.2 | 7 | 1971 | 20.5 | 6 |
| Canada | 1970 | 78.1 | 2 | 1970 | 28.5 | 2 |
| Denmark | 1970 | 51.7 | 9 | 1972 | 23.7 | 3 |
| Finland | 1967 | 47.4 | 10 | 1972 | 13.8 | 13 |
| France | 1970 | 54.3 | 6 | 1972 | 17.0 | 10 |
| Germany (FDR) | 1969 | 30.5 | 16 | 1972 | 17.1 | 9 |
| Ireland | 1971 | 47.0 | 11 | 1970 | 12.2 | 14 |
| Italy | 1966 | 30.8 | 15 | 1972 | 20.1 | 7 |
| Luxembourg | 1970 | 37.4 | 13 | 1972 | 1.9 | 16 |
| Netherlands | 1970 | 52.5 | 8 | 1971 | 21.2 | 5 |
| Norway | 1970 | 68.9 | 3 | 1972 | 19.0 | 8 |
| Sweden | 1972 | 68.1 | 4 | 1972 | 22.4 | 4 |
| Switzerland | 1970 | 62.6 | 5 | 1972 | 10.3 | 15 |
| United Kingdom | 1970 | 39.4 | 12 | 1972 | 15.0 | 11 |
| United States | 1970 | 82.9 | 1 | 1972 | 51.5 | 1 |

*Source:* Rates for 15- to 18-year-olds adapted from UNESCO (1974, Tables 15 and 16); rates for 20- to 24-year-olds adapted from UNESCO (1975, Table 3-2).

The United States enrolls a higher percentage of young people in secondary school and college than does any other Western nation.

Torsten Husen, professor in the Institute of International Education at the University of Stockholm and an expert in the international assessment of school performance, explains why international comparisons do not accurately assess the performance of the U.S. school system. All comparisons must begin from a base of similarity, or they fail, as in the proverbial comparison of apples and oranges. The differences between American and European schools are so great that a base of comparison is nearly impossible to establish. Even general terms like "primary" and "secondary" education undergo transformation when they cross national boundaries.

European primary and secondary schools differ from those in the United States in governance, purpose, and enrollment. In West Germany, for instance, both primary and secondary schools are under the control of a central state or land government (like Bavaria). Primary schools are at least partially funded by the government, and secondary schools are wholly so. All German schools are inspected by representatives of a central agency, usually a state ministry of education. At the level of the secondry school, the difference in purpose is even more striking to an American than the difference in governance. European secondary schools have traditionaly existed for students who from their earliest years attend private rather than public institutions, and these public secondary schools still exist to educate an elite and to train them for entrance to the universities. In England and Germany, the last year of secondry school includes whose who in the United States would be in the college freshman age group.

The attack on the elitism of the European systems during the 1960s had its effects: students are no longer tracked as early or as consistently as they previously were, although by our standards European secondary schools are still elitist. Before the 1960s, students entering academic secondry schools

transferred into them from grade 4 or 5 of primary school. Secondary school occupied six to nine years and covered a curriculum geared to requirements set by the Universities. In the 1950s, only 5 to 10 percent of the pupils in Western and Northern Europe completed academic secondary school and took the "school-leaving" examination that also served as an entrance examination to the Universities. From the mid-fifties to the early seventies, enrollment in European secondary schools greatly increased but still included only about 20 percent of the relevant age group. In some countries, the percentage was even less. In the mid-sixties, when 75 percent of the pupils who had entered the system in the first grade were graduated from secondary school in the United States, only 9 percent were graduated in West Germany.

## U. S. versus West German Schools

To compare American schools with West German schools, then, is to compare a compresensive system with an elite system. At no point are the two similar enough to bear comparison. To compare secondary school achievement in Germany and in the United States would be to measure a university-bound group in Germany against a general youth population in the United States. Even if one compared the University-bound groups in the two countries, a base of comparison would be lacking. In Germany, the university-bound group represents a much smaller percentage of the youth population than the corrresponding group in the United States, and its members have been selected much earlier.

Furthermore, the curriculum of the two systems varies greatly. In the United States, even the elite group--in terms of academic achievenment--has been prepared in a comprehensive curriculum. A large portion of their time in school has been spent in the study of subjects for which there is no parallel in West German schools. There is no large vocational track in European secondary schools; vocataional training is normally conducted at the employer's workplace. Nor do students study health or physical education or crafts. When I visited West Germany, I was struck with the brevity of their school day--about 8:00 a. m. to 12:30 p. m. Because homework is assigned a much higher priority, school hours do not need to be as long as ours, and all extracurricular activities are

conducted by community agencies rather than by the schools.

## Comparing Elite and Comprehensive Systems

When basic differences in structure and enrollment between our schools and those of other industrialized countries are taken into account, we see that our schools are performing better than the headlines would lead us to believe. Our high school graduates score higher on international assessment tests in reading than graduates in all but three other Western industrial nations, despite the fact that those countries graduate a much smaller percentage of their students. When adjustments are made for the great difference in the percentage of young people tested, our students also compare favorably in mathematics and science (see Figures 1 and 2).[4]

Husen explains how to adjust for these differences: Compare only the small percentage at the top level of the achievement band. Each system presumably retains the most able students, but while selective systems retain only those, comprehensive systems retain them as an elite within a larger group. Given that selective systems value most highly the quality of the end product--the elite--to compare the elite of a comprehensive system with the end product of a selective system is fair to both systems.

To illustrate: In the United States, high school seniors include about 75 percent of 17- to 18-year-olds. In Sweden, students who finish grades 10 to 12 (gymnasium) make up about 50 percent of the 17- to 18-year-old age group. In the Federal Republic of Germany, grade 13 (oberprimaner) accounts for only about 15 percent of the same age group. When tests compared the average performance of students in the terminal grade studying mathematics (see Table 1), American high school seniors placed far below "seniors" in other countries. But in the United States, 18 percent of 17- to 18-year-olds took mathematics as against 4 to 5 percent in some European countries. To make a more accurate comparative assessment, the testers took this smallest percentage as a base. When the comparison was based on the average performance of the top 4 percent of the

relevant age group across all participating countries, the range of scores was much narrower (see Figures 1 and 2).

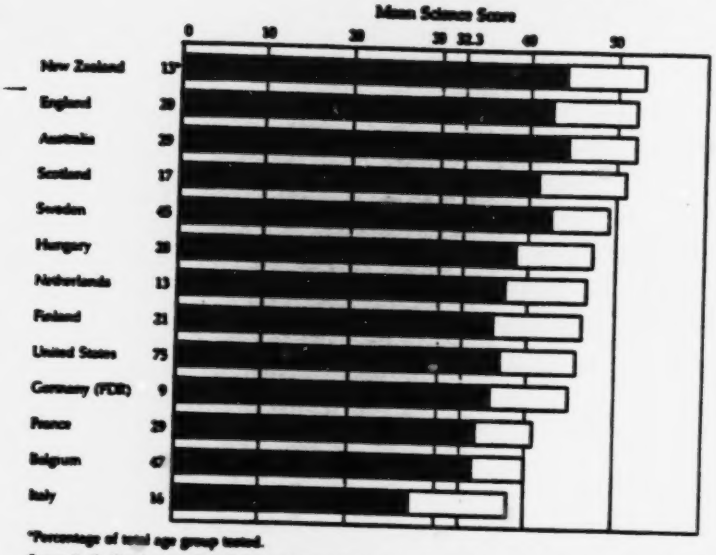
Husen concludes,

> the international surveys of both mathematics and science demonstrated that the top 5 percent to 10 percent at the end of secondary education (i.e., the elite) tended to perform at nearly the same level in both comprehensive and selective systems of secondary education. Thus the elite among U. S. high school seniors did not differ considerably in performance from their age-mates in France, England, or Germany. The comprehensive systems, where the net is cast more widely, result in a bigger talent catch. In addition, those who are less able get a better opportunity to develop their potential than in the selective systems of the traditional European type.[2]

Surveys of international achievement teach another useful lesson: Opportunity to learn is the single factor that best explains differences in performance among students of different countries. A study of French as a foreign language compared eight countries in terms of time spent on the subject (number of years of study); the grade level at which French was introduced in school; and the number of periods of instruction per week. Romania, which requires students to study French for about six years, scored at the top; the United States, which rarely requires French and when it does, requires two years of study at the most, scored at the bottom.

FIGURE -2

**An International Comparison of Test Scores in Science of Students in Last Year of Secondary School**




*Percentage of total age group tested.

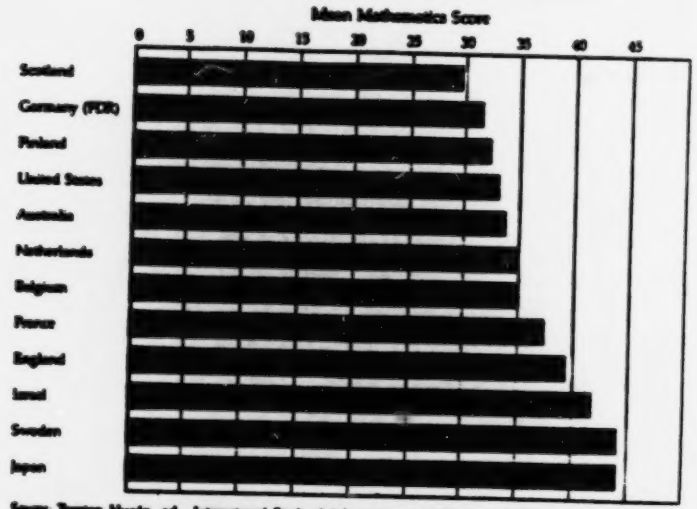
Source: L. C. Comber and John P. Keeves, *Science Education in Nineteen Countries* (Stockholm and New York: Almquist and Wiksell and Wiley-Halsted Press, 1973).

■ Overall Score ■ Top 9% ■ Top 5% □ Top 1%

International comparisons involving the United States are skewed by the egalitarianism of the U.S. system. When scores compared are adjusted to reflect comparable proportions of the relevant age group, U.S. scores are similar to those of other countries.

FIGURE -1

**An International Comparison of Test Scores in Mathematics of Students in Last Year of Secondary School**




Source: Torsten Husén, ed., *International Study of Achievement in Mathematics: A Comparison Between Twelve Countries* (New York: Wiley, 1967).

■ Average Mathematics Test Score for Equal Proportion of Age Groups
■ Average Mathematics Test Score for Country

When scores of all students studying mathematics in the terminal year of secondary school are compared, the United States ranks last, in part because U.S. schools enroll a much higher percentage of the relevant age group in mathematics (18 percent as against 5 percent in some European countries). When scores of equal proportions of the age group are compared, U.S. students perform about as well as students in Australia, Belgium, and West Germany.

Husen also calls attention to negative trends in European education and documents wave of criticism. Some critics complain of low standards and lack of discipline, while others accuse the schools of being joyless, oppressive, and autocratic. In Europe, as in this country, national commissions are criticizing the secondary schools. There, too, educational institutions have been accused of failing to meet their own goals of society's needs. In the Netherlands and in the Federal Republic of Germany, as in this country, studies show that as they progress through the compulsory schooling grades, from ages 5 through 14, many students become increasingly negative about school. Similarly, too, in the 1960s, when expenditures for education across the Western world grew twice as fast as the gross national product, demands were raised for accountability and tighter public control. Then, during the 1970s, in Germany, Great Britain, and Australia, as in the United States, education slipped down the scale of national priorities, budget growth rates slowed down, and school people practiced the "management of decline." As Table 2 demonstrates, the U.S. effort in public school expenditures compares favorably with others nations'.

## U.S. versus Japanese Schools

If the comprehensive nature of our school system makes comparison between U.S. students and West German students difficult, there is another country, Japan, where the elementary/secondary schools are at least as egalitarian as our own. (The United States, however, still provides more opportunity for students

TABLE -2

**An International Comparison of Expenditures for Public Schools**

| Country | Year | Expenditure (in millions) | Percent of GNP | Percent of Total Public Expenditure |
|---|---|---|---|---|
| Sweden | 1978 | 7,850 | 9.1 | 13.1 |
| Canada | 1978 | 16,153 | 8.1 | 18.5 |
| Saudi Arabia | 1979 | 4,841 | 7.4 | 10.2 |
| Soviet Union | 1978 | 47,431 | 7.4 | 12.0 |
| Australia | 1977 | 6,402 | 6.5 | 16.2 |
| United Kingdom | 1976 | 14,177 | 6.3 | 14.3 |
| United States | 1977 | 120,700 | 6.3 | 17.7 |
| Austria | 1978 | 3,280 | 5.7 | 8.0 |
| Japan | 1978 | 55,563 | 5.7 | 16.1 |
| Brazil | 1979 | 1,043 | 5.6 | (NA) |
| France | 1978 | 25,301 | 5.3 | (NA) |
| Mexico | 1979 | 6,103 | 5.2 | (NA) |
| Zaire | 1979 | 363 | 5.0 | 25.8 |
| Germany (FDR) | 1979 | 34,699 | 4.6 | (NA) |
| India | 1977 | 2,923 | 2.9 | 9.9 |
| Columbia | 1978 | 513 | 2.3 | 19.9 |

Source: U.S. Bureau of Census, *Statistical Abstract of the United States*, 1983.

U.S. expenditures for public education compare favorably with those of other countries.

to acquire a postsecondary education.) In Japan, some 90 percent of all children graduate from the twelfth grade. Furthermore, the Japanese schools have raised the whole base of their youth to a higher level of understanding in mathematics and science. All Japan's high school graduates must complete at least two years of math and two years of science, whereas U.S. school districts typically require only one year of each. College-bound Japanese students study math every year high school and attain a level of sophistication beyond trigonometry. In contrast, only 11 percent of high school students in California reach trigonometry. The college-bound Japanese generally take physics, chemistry, biology, and earth sciences, while in the U.S. many college-bound students assiduously and successfully avoid any such array of science courses.

This emphasis continues in college, with the result that Japan graduates more engineers from undergraduate college programs than does the U.S., despite the fact that their population is only half the size of ours. The U.S. produces Ph.D. researchers in numbers large enough to compete with Japan, but production of technicians in Japan outstrips that of the U.S. by a considerable margin. The Japanese are setting very high standards of scientific literacy for their entire school population and at the same time developing an intensive science and math track for their ablest students.

Japan also requires all of its high school students to take an extensive language and social studies curriculum, including ethics, civics, history, political science, and economics. All high school graduates must complete at least three years of social studies, whereas U.S. school districts generally require only two years. While Japan was increasing its social studies enrollments in the 1970s, the percentage of California high schools students taking a social studies course in grades 7 through 12 declined from 71 percent to 45 percent - a drop of 575,000 students in one state alone.

The overall achievement scores of Japanese students in math and science are the highest in the non-Communist world. And Japanese pupils learn more social studies than do our students. How are they able to accomplish so much? One

TABLE 3

**A Comparison of Instructional Time in Nine Countries**

| | Instructional Hours per Day | | Instructional Days per Year | | Instructional Hours per Year | |
|---|---|---|---|---|---|---|
| | Primary | Secondary | Primary | Secondary | Primary | Secondary |
| Belgium | 4.2 | 5.6 | 200 | 200 | 840 | 1,120 |
| Finland | 4.0 | 6.0 | 200 | 194 | 800 | 1,164 |
| France | 5.4 | 5.0 | 170 | 155 | 918 | 775 |
| Germany (FDR) | 3.3 | 4.0 | 230 | 230 | 759 | 920 |
| Italy | 5.0 | 5.8 | 175 | 195 | 875 | 1,131 |
| Netherlands | 5.2 | 5.6 | 200 | 200 | 1,040 | 1,120 |
| Sweden | 5.4 | 6.0 | 180 | 180 | 972 | 1,080 |
| United Kingdom | 5.0 | 5.0 | 180 | 180 | 900 | 900 |
| United States | 5.0 | 5.0 | 180 | 180 | 900 | 900 |

Source: A. Harry Passow et al., *The National Case Study: An Empirical Comparative Study of Twenty-one Educational Systems* (New York: Wiley and Sons, 1976).

answer is time. The level of student achievement is directly related to the amount of time spent studying specific curricular content. If Group A spends twice as much time studying algebra as a similar group of students, Group B, then on the average Group A will show higher math achievement than Group B. Educational researchers use a specific phrase, academic learning time, to refer to the amounts of time that a student spends in learning tasks that yield high rates of success. Japanese schools spend much more time on academic tasks than do their U.S. counterparts. Their school year is longer by about one-third, and the Japanese high schools offer only a fraction of the non-academic electives that occupy - or clutter - the curriculum in the United States. If we compare the U.S. and Europe we find that the total time spent in elementary school is comparable. But when we get to the high school level, several European nations provide more instructional hours per year than we do (see Table 3).

Japanese students also spend far more time doing homework than do American students -- nearly twice as many hours as their counterparts in California, for example. Japanese students know that their chances for economic success depend on being accepted at a good college. Few Japanese ever change employers after college, and the best jobs with the best companies go to graduates of the most competitive colleges. Universities in Japan select students according to high school grades and the results of rigorous college entrance exams. Thirty-nine

percent of Japanese high school graduates go on to postsecondary education soon after high school, compared with 44 percent in the United States.

## Cultural Differences and School Performance

All these facts illustrate another truth: A comparison of student achievement and school performance is also a comparison of societies. In internationally based tests we have no way to adjust for such related variables as the homogeneity of the society and the social and economic condition of its members. For example, some researchers looking for the sources of the successful performance of Japan's school system have found a familiar figure --mother. Jerome Kagan, a psychologist at Harvard University, describes a typical family pattern:

> Until her child goes to school, the Japanese mother devotes herself to the rearing of the child. In verbal and nonverbal ways, she reminds the child of her deep, deep, warm feelings and that the child is the most important thing in the world to her. Then she says, "After all I've done for you, don't disappoint me." She's like the Jewish mother who says, "What do you mean you're not hungry - after I've slaved all day over a hot stove for you?"[3]

While we have chosen to emphasize the values of equality and individualism and thus the right of women to work outside the home, Japan tracks most of its women into the equivalent of the now passe' American "finishing school," where the emphasis is on traditional homemaking skills.

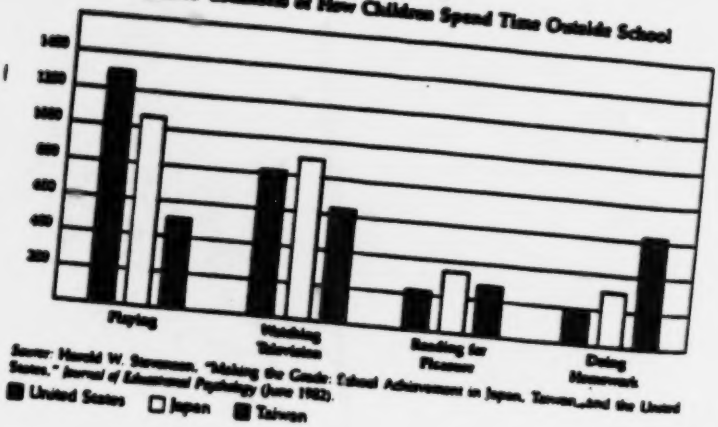
Cultural differences also appear to affect differences in educational achievement within our own society. In 1975, the Educational Testing Service, which develops the aptitude test used by many colleges to screen applicants, began to record test scores by ethnic group. The results showed that Asian-Americans, mainly Chinese, score about 45 points above the national average on the Scholastic Aptitude Test -- higher than any other ethnic group. On the Graduate Record Exam, used to screen college graduates applying for advanced study, the same ethnic group averages about 55 points above the national norm in mathematics. Why? The New York Times of January 9, 1983,

reports some answers. The dean of Cooper Union's School of Engineering, Dr. Chor-Wen Tan, cites love of education and a gift for detail and precision. He then says, "American children are trained to ask the question 'Why?' but Chinese children are taught to listen and to quietly reflect." A professor of mathematics and education at Pennsylvania State University emphasizes memorization--the traditional method of learning in China and a discipline useful in the study of mathematics. The chairman of Wang Laboratories in Massachusetts, Mr. An Wang, who was born and raised in Taiwan, suggests that the language itself may make mathematics seem more natural to Chinese: "In the Chinese language there is the one-syllable character, and Arabic numbers are also one syllable. Chinese may simply find it easier to read numbers than words. "But the productivity of students in the United States to ask "Why?" and to question traditional ways of doing things contributes to the innovative and entrepreneurial character of our nation. Each of these societal characteristics has benefits and costs.

Another international study also stresses cultural differences underlying differences in school performance. A study organized by the Center for Human Growth and Development at the University of Michigan and directed by Harold W. Stevenson compared achievement, behavior, and parental (maternal) attitudes for 5,000 young pupils--first-graders and fifth-graders--in 120 classrooms in each of three cities: Minneapolis, Minnesota; Taipei, Taiwan; and Sendia, Japan. At both grade levels, the American children scored poorly in mathematics. If children from the three cities had performed comparable, approximately 33 children from each country would have been among those receiving the top 100 scores. For American first-graders, the number was 15. (Sixteen percent of the first-graders could not count the 17 dots in one of the test questions.) For American fifth-graders, the number was one. In none of the 20 American fifth-grade classrooms studies did the children reach an average score on the mathematics test equivalent to that of the children in the worst-performing Japanese classroom. (Ten percent of the American fifth-graders could not divide 42 by 6, and 9 percent could not divide 24 by 3.) Although the American pupils scored better in reading than in math, the test scores for reading and

FIGURE 3

Mothers' Estimates of How Children Spend Time Outside School

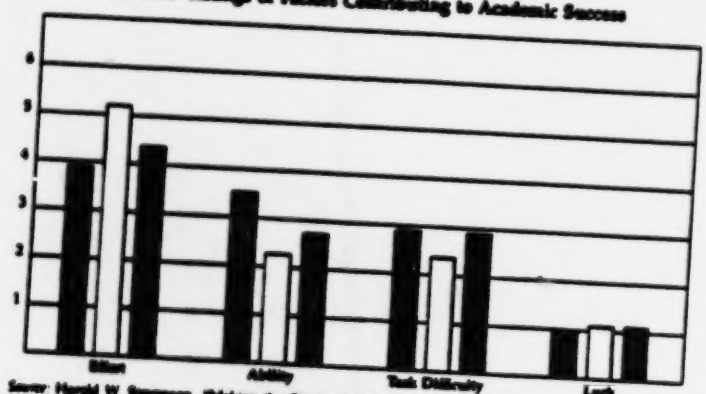



Source: Harold W. Stevenson, "Making the Grade: School Achievement in Japan, Taiwan, and the United States," *Journal of Educational Psychology* (June 1982).

■ United States □ Japan ■ Taiwan

FIGURE 4

Mothers' Ratings of Factors Contributing to Academic Success




Source: Harold W. Stevenson, "Making the Grade: School Achievement in Japan, Taiwan, and the United States," *Journal of Educational Psychology* (June 1982).

■ United States □ Japan ■ Taiwan

vocabulary showed the American pupils over-represented among both the best and the worst readers.

Stevenson and his fellow observers also found that American children were far less attentive to the teacher and to their work than were the Japanese or Chinese children. They were more inclined to talk to their classmates, to ask irrelevant questions, to wander about the room, or simply to stare into space. At home, the Americans spent much less time on homework. They also watched more television and did more chores (see Figure 3). One Chinese mother, when asked why she did not assign her child any chores, responded, "It would break my heart to take him away from his studies." American mothers more frequently reported themselves satisfied with their children's progress in school than did the Chinese or Japanese mothers. American mothers were also more likely than their counterparts to credit school achievement to the child's natural ability. The Chinese--and especially the Japanese -- mothers placed more emphasis upon the child's efforts (see Figure 4).

## Is Japan Our Best Model?

Even in Japan, however, educators acknowledge problems, and in some areas they look to our system for models. For example, Japan's math and science curriculum makes widespread use of instructional materials developed in the United States. It is ironic that the U.S. National Science Foundation developed

the new physics, chemistry, and biology materials that the Japanese have adapted. These curricular materials are among the most challenging available, and Japan makes an enormous investment in keeping science teachers up-to-date. By contrast, the U.S. does very little in-service teacher training and is experiencing shortages of math and science teachers.

My direct observation of classroom teaching techniques in Japan reveals some weaknesses that the United States may be able to avoid. A persistent teaching strategy in Japan is the use of imitation and rote learning--methods considered outmoded by most educators in this country. In art classes, for example, students laboriously copy ancient ceramic treasures.

Japanese high school students are judged on standardized tests primarily by their memorization of facts and concepts. Japanese teachers dispense knowledge, and the students dutifully write it down as unchallengeable truth. They do not use the school library creatively to weave together sources of information from which to formulate their own interpretations. This style of education is not noted for enhancing innovation or for fostering radical departures from accepted ways of thinking. On the other hand, Japanese children are taught that each repetition of a process contains something new. They learn to discriminate among tiny variations in routines, to "see the form, but see through the form to improve it." The Japanese record of improvement of imported technologies attests to the success of this philosophy.

Many in our country are more than a little envious of Japanese economic success and admire the schooling that enhances it. But we need not rush to copy the Japanese system. The educational achievements of that nation have been won at a high price. A major concern of the Japanese Ministry of Education is that academic competition is too intense. Officials are worried about the extent of psychological stress associated with the highly competitive college entrance examinations.

They fear that Japanese children are neglecting other aspects of their development, among them physical fitness. Getting into a good college is one of

the few legitimate goals that Japanese society offers its young people (some say the only goal). Intense academic pressure also produces negative social effects. Not surprisingly, some nonachievers lose all sense of meaning in their lives and turn to truancy, gangs, and even suicide, partly because non-academic vocations command little prestige.

In assessing what these international tests tell us about how well schools perform in Japan or West Germany or any other country, we should remember the limitations of all tests. Tests are biased by the purposes for which they are designed. Some tests used in international comparisons were designed exclusively to select the most able students for entrance into a university. They were not designed to assess how well average students handle the material they have studied, nor were they meant to constitute an international School Achievement Olympics. In short, they do not perfectly match the schools curricula, nor do they test how well the schools are performing the roles assigned to them in their respective societies.

Furthermore, tests can evaluate some areas of skill much better than others. To construct a test that would show how well the schools are doing in transmitting a common culture or how well students have learned their country's literature or history would call for remarkable feats of test engineering. Do French secondary schools do a better job of teaching Racine or Moliere than the secondary schools in our country do of teaching Shakespeare? The answer would depend to some extent on the relative difficulty of the language of each writer. Which Japanese novelists compare best with Dickens or with Melville?

An analogy here from the world of sports illustrates the difficulty: Swimming and gymnastics are practiced in the same way throughout the world, but "football" means, variously, rugby, or soccer, or the activities of the NFL. Mathematics and science have the same characteristics wherever they are found, but literature varies in dozens of ways. Even "simple" assessments of reading are difficult to make. How can a test adjust for the varying degree of difficulty among national languages? If English comprises a greater total number of words than German or Japanese, is the task of learning to read it more

difficult? How great a barrier to development of reading skills is the notorious phonetic irregularity of English spelling? Given the difficulty of creating an appropriate instrument to test students' skill in reading literature or history even in their native language, it is not surprising that international tests focus upon math and science. And because Japanese schools also focus on math and science, the tests are skewed toward the greatest strength of the Japanese curriculum. From this perspective, it is less surprising that Japanese test scores in these areas are outstanding. The international tests, then, show us a few things, but not what the newspapers sometimes report them as showing.

TABLE 4

**Scholastic Aptitude Test Score Averages for College-Bound High School Seniors, 1966–81**

| | 1966-67 | 1970-71 | 1975-76 | 1978-79 | 1980-81 |
|---|---|---|---|---|---|
| **Average Verbal Scores** | | | | | |
| Male | 463 | 454 | 433 | 431 | 430 |
| Female | 468 | 457 | 430 | 423 | 418 |
| Total | 466 | 455 | 431 | 427 | 424 |
| **Average Mathematical Scores** | | | | | |
| Male | 514 | 507 | 497 | 493 | 492 |
| Female | 467 | 466 | 446 | 443 | 443 |
| Total | 492 | 488 | 472 | 467 | 466 |

*Source:* Department of Education, National Center for Education Statistics, *Digest of Education Statistics* (Washington, D.C.: U.S. Government Printing Office, 1980, 1982); College Entrance Examination Board, *National Report, College Bound Seniors* (New York: CEEB, 1981).

The SAT, designed to assess a pupil's potential for college-level work, does not directly reflect what students have been taught in elementary and secondary school. Declining scores (now stabilized) have been news for a decade, even though the test does not accurately reflect the performance of the schools.

## Declining Test Scores

But what about measurements made by our own tests, measurements that reveal a decline in students' scores over the past ten years? Reports on the declining scores of graduating high school seniors on the Scholastic Aptitude Test (SAT) have made news for a decade. As Table 4 indicates, there was a decline between 1966 and 1980, but this decline was arrested in 1981 and scores are now stable. The most recent tests (1984) even show some improvement. But the SAT tests were never meant to be a thermometer for the entire system of education in the United States. The SAT is designed to assess a pupil's potential or innate aptitude for college-level work, but it does not reflect what students have been taught in elementary and secondary schools. Also, because it tests only the college-bound population of the schools, it does not measure the performance of students at the low end of the band of academic achievement. An immense number of such students could have improved their skills over the same period of time, but the SAT would record nothing of this achievement.

A more comprehensive study of performance is the National Assessment of Educational Progress (NAEP). Initiated in 1969, the NAEP is funded by the U.S. government and administered by the Educational Testing Service in Princeton, New Jersey. The NAEP includes all students, not just those who are college bound, and it measures what students have actually learned. The NAEP shows that

between 1973 and 1980, the average score in mathematics earned by 17-year-olds declined 3.6 percentage points. On the other hand, from 1974 to 1980, the reading scores of 9-year-olds increased by 2.6 percentage points. The NAEP, then, agrees with the SAT in recording a decline in mathematical skill among seniors in secondary school, but it also records the schools' success with younger pupils.

NAEP results do, however, emphasize the growing weakness in U.S. students' attainment of "higher order skills." U.S. youngsters can comprehend short paragraphs but have great difficulty with longer, complex passages, particularly with respect to interpreting the tone and mood of the author. U.S. children do well in repetitive math problems like multiplication tables but are losing ground in math problem solving. This includes the applciation of mathematical formulas or concepts to problems or contexts that the student has not been drilled on or seen before. (Large gains were made by black pupils compared to white since 1970, but there is still a considerable gap to close.)

In short, the age group more fully under the influence of the school, the 9-yar-olds, improve in a key academic skill, while that group most likely to be influenced by peers, to abuse drugs and alcohol, and to cut class -- the teenagers - has lost ground in mathematics. This way of interpreting the facts is obviously biased, but it is meant to be -- to illustrate how little these tests take into account social and cultural conditions.

Between 1968 and 1974 the NAEP reported a considerable drop in U.S. science achievement scores -- and it has not been reversed in later tests. Ralph Tyler, former director of the Center for Advanced Study in the Behavioral Sciences at Stanford, argued that there is a correspondence between public attitudes toward science and the average level of science achievement among 17-year-olds. According to this view, negative public attitudes about science turned many students away from science. At the time of the first national assessment in 1969, the U.S. public was enthusiastic about our space program and its technological achievements. But by 1974, when scores dropped, reports of air and water pollution, of excessive exploitation of natural resources, and of the

harmful effects of technology had brought about a decline in the prestige of science.

The NAEP does, however, document a significant achievement for American education... The scores for black students registered gains in achievement, especially for students with low scores in math. This group showed a substantial improvement between 1972 and 1980 in elementary and middle schools. Thus, although 13-year-olds in general made lower scores on the mathematical assessment in 1980 than in 1973, black students raised their scores over the same period, with low achievers showing greater gains than high achievers.

This gain was not spontaneous or accidental. During the 1970s, the federal and state governments placed a high priority on improving the achievement of disadvantaged students, particularly Blacks in the Southeast. Title I of the federal Elementary and Seondary School Act provided compensatory education programs such as remendial reading and math. Many of these programs employed specialist teachers who worked intensively with small groups of pupils. The nation persisted in this objective from 1965 to 1980, and by 1972, when the Title I program was seven years old, the major positive results were already manifest. By the 1980s, the achievement tests measured a significant gain for the group aided through Title I. The lesson is clear: When we define our goals and pursue them -- with patience -- we can improve our schools' performance

## Evaluating the Tests

Both the international tests and the NAEP tests assess only the most academic, the most intellectual, areas of learning. International testing leaves many questions unanswered. Do American pupils have higher self-esteem, better citizenship skills, or more creativity? Do our students know more about potential career choices and enjoy more enriching leisure activities? How do our young musicians and artists fare in international competitions? The international tests cannot tell us. Have courses in driver's education lowered the number of arrests for reckless driving or reduced the number of accidents? Does school improve our students' health and eating habits? When our pupils

become adults do they spend more of their leisure time in reading? The international tests cannot tell us. Nor can the national tests. All of these questions focus on goals that are important for sections of the American public, but we do not assess whether the schools are meeting them. To boil down the performance rating of our schools to SAT measure of college aptitude for one-third of the graduating high school students is ridiculous. A headline in the Los Angeles Times of April 15, 1984, sums up the situation: "U.S. Schools Lack a Yardstick to Measure Performance -- A Sticky Situation."

Nor does it make sense to believe that international tests have revealed so great a weakness in our schools that it threatens our economic standing in the world. The catalyst of an earlier crisis -- the orbiting of Spuntnik -- revealed, among other things, that President Eisenhower and Defense Secretary Wilson had chosen not to accelerate our manned space program in order to concentrate on other uses of rockets. But the schools ended up among the main villains when the Soviet Spuntnik launch beat the U.S. timetable. Similarly, Japan's emergence as an economic threat to the U.S. can be trace to the history of the American steel and auto industries as well as to the American school system. To expect the school system to solve problems of rigidity and low quality in our industrial economy is once again to expect public education to solve every social problem from alcoholism to unemployment.

The object of international testing is not to create a quasi-Olympics but to work toward the design of instruments to measure school performance. If we pull back from a competitive and crisis-ridden view of the results of these tests, we can learn much that is useful. For example, we can learn that when adjustments are made for our larger enrollments, schools in Europe do not appear to prepare students better than we do. We learn, too, that our best students compare favorably with the best in any country. And, perhaps most important of all, we learn that the only other nation to have over 75 percent of its pupils enrolled in high school has a record of high academic achievement. That an egalitarian system like Japan's reaches a high level of academic achievement ought to be goods news, given the importance of equality in our system of valuves.

It is not good news, however, to learn that the gap in academic educational achievement between Japan and our nation is increasing. But we know that the goals of Japanese schooling do not include vocational or personal development. Japanese schools limit themselves to academic objectives, and they commit more time to those objectives. Their school year is 240 days, as compared to the U.S. school year of 175 days. West German schools also limit their objectives and seem to accomplish these goals even though their school day is shorter than ours. From these comparisons and from the success of our Title I programs of the 1970s, we can draw encouraging conclusions that support one of the major themes of this book: If we reduce the number of objectives that our schools pursue, and if we show patience and persistence as the schools pursue those goals, we can improve the performance of the public schools and raise the level of our students' academic achievement.

## Chapter 7

### NOTES

1) Horace Mann, Twelfth Annual Report of the Secretary of the Board of Education, West Newton, Massachusetts, November 24, 1848 (Boston: Dutton, Wentworth State Printers, 1849).

2) Torsten Husen, "Are Standards in U.S. Schools Really Lagging Behind Those in Other Countries?" Phi Delta Kappan (March 1983) p. 458.

3) Perry Garfinkel, "The Best Jewish Mother in the World," Psychology Today (September 1983) p. 56.